2018-06

APPLICATION OF BAYESIAN STATISTICAL POST-PROCESSING TECHNIQUES TO PROBABILISTIC NOWCASTS OF CEILING HEIGHT AND VISIBILITY

Jones, Kellen T.

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

APPLICATION OF BAYESIAN STATISTICAL POST-PROCESSING TECHNIQUES TO PROBABILISTIC NOWCASTS OF CEILING HEIGHT AND VISIBILITY

by

Kellen T. Jones

June 2018

Thesis Advisor: Wendell A. Nuss
Second Reader: Joel W. Feldmeier

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2018	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
APPLICATION OF BAYESIAN STATISTICAL POST-PROCESSING TECHNIQUES TO PROBABILISTIC NOWCASTS OF CEILING HEIGHT AND VISIBILITY

5. FUNDING NUMBERS

6. AUTHOR(S) Kellen T. Jones

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

Nowcasting is a modern technique in weather prediction that seeks to produce highly accurate analysis and short-term forecasts of up to six hours. Challenges to nowcasting include numerical forecasting spatial and temporal resolution and data availability, especially in data-denied or limited regions. Nowcasting cloud ceiling height and horizontal visibility is a specific example of a challenging nowcasting problem.

A nowcast system is applied and tested on summertime conditions from June to August 2017 over the Monterey Regional Airport in California. The system post-processes 12 km North American Mesoscale Model (NAM) data from a local grid point to produce short-term multivariate probabilistic predictions of ceiling of height and visibility. Bayesian Estimation (BE) and Monte Carlo Markov Chain (MCMC) methods are used to train the system from a set of past predictor variables and observations.

The approach demonstrates error reduction and skill improvement over the raw NAM ceiling height and visibility forecasts. The computationally cheap system also explicitly communicates uncertainty and requires a relatively limited training data set compared to other statistical post-processing techniques. Using short training periods and/or analog techniques, this system can be used to nowcast in regions with limited or no observational data availability.

14. SUBJECT TERMS operational nowcasting, cloud forecasting, Bayesian estimation, statistical post-processing, supervised machine learning, ceiling height, visibility, probabilistic weather forecasting	**15. NUMBER OF PAGES** 95
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

APPLICATION OF BAYESIAN STATISTICAL POST-PROCESSING TECHNIQUES TO PROBABILISTIC NOWCASTS OF CEILING HEIGHT AND VISIBILITY

Kellen T. Jones
Lieutenant, United States Navy
BA, University of California - Los Angeles, 2010

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN METEOROLOGY AND PHYSICAL OCEANOGRAPHY

from the

NAVAL POSTGRADUATE SCHOOL
June 2018

Approved by: Wendell A. Nuss
Advisor

Joel W. Feldmeier
Second Reader

Wendell A. Nuss
Chair, Department of Meteorology

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Nowcasting is a modern technique in weather prediction that seeks to produce highly accurate analysis and short-term forecasts of up to six hours. Challenges to nowcasting include numerical forecasting spatial and temporal resolution and data availability, especially in data-denied or limited regions. Nowcasting cloud ceiling height and horizontal visibility is a specific example of a challenging nowcasting problem.

A nowcast system is applied and tested on summertime conditions from June to August 2017 over the Monterey Regional Airport in California. The system post-processes 12 km North American Mesoscale Model (NAM) data from a local grid point to produce short-term multivariate probabilistic predictions of ceiling of height and visibility. Bayesian Estimation (BE) and Monte Carlo Markov Chain (MCMC) methods are used to train the system from a set of past predictor variables and observations.

The approach demonstrates error reduction and skill improvement over the raw NAM ceiling height and visibility forecasts. The computationally cheap system also explicitly communicates uncertainty and requires a relatively limited training data set compared to other statistical post-processing techniques. Using short training periods and/or analog techniques, this system can be used to nowcast in regions with limited or no observational data availability.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

AGL	above ground level
AMS	American Meteorological Society
ASOS	Automated Surface Observing System
BE	Bayesian estimation
BMA	Bayesian model averaging
CDF	cumulative density function
CLG	ceiling
DTG	date time group
FAA	Federal Aviation Administration
FAR	false alarm ratio
GLM	generalized linear model
GRIB	gridded binary files
HR	hit ratio
HSS	Heidke skill score
IFR	instrument flight rules
IQR	interquartile range
JCS	Joint Chiefs of Staff
KDE	kernel density estimation
KMRY	Monterey Regional Airport
LAMP	Localized Aviation MOS Program
LIFR	low instrument flight rules
MAE	mean absolute error
MCMC	Markov Chain Monte Carlo methods
ME	mean error
MOS	model output statistics
MSL	mean sea level
MVFR	marginal visual flight rules
NAM	North American Mesoscale model
NCAR	National Center for Atmospheric Research

NCEP	National Center for Environmental Prediction
NPS	Naval Postgraduate School
NRC	National Research Council
NWP	numerical weather prediction
NWS	National Weather Service
PC	percent correct
PDF	probability density function
PPD	posterior predictive distribution
RMSE	root mean square error
SREF	Short Range Ensemble Forecast
SS	skill score
TAF	terminal aerodrome forecast
TS	threat score
UAV	unmanned aerial vehicle
VFR	visual flight rules
VIS	visibility
WFO	Weather Forecast Office

ACKNOWLEDGMENTS

This thesis is dedicated to my wife, Diana, and son, Christopher. They give more meaning to my work and mission in life than I could have ever imagined. Their support and patience during this experience has been crucial to my success and continued commitment as a naval officer and lifelong student. I love you both more than you will ever know.

I would like to thank my parents Tim and Gloria for always encouraging my search for truth and reason in life. My love of science and the wonders of the natural world are founded on the love of nature my mother passed down to me. My father taught me that I can do anything if I only follow my heart. Thank you both. To my brother Tim and his family, thank you for always encouraging me and making me a proud uncle.

Finally, I would like to thank all my mentors and colleagues that have carried me to this point. To Rear Admiral (Ret) Tim Gallaudet and Dr. Bill Burnett, thank you for your mentorship and intellectual encouragement. To Chief Yeoman Derrick Pickens, thank you for your advice and support. There is no Navy without the Chief. To my advisors, Dr. Wendell Nuss and CDR Joel Feldmeier, thank you both for giving me the freedom to explore this thesis and for your sage advice. To Commander Paula Travis, thank you for your leadership and for encouraging me to confidently get out of my swim lane. To Lt. Commander Travis Wendt, thank you for your mentorship, support, and for entrusting me as the curator of your work. To Bob Creasey, Mike Cook, and Mary Jordan, your technical expertise is unmatched and critical to this thesis. To my cohort, thank you for making the past 2.5 years a blast and an example of true teamwork and camaraderie.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. CONTEXT

In order to make the best strategic, operational, and tactical decisions, military commanders must maintain superior battlespace awareness and prevent the adversary from gaining a competitive advantage in awareness. To that end, the United States' competitors are currently accelerating their ability to acquire and use information in warfare. According to the Chief of Naval Operations, Admiral John Richardson, our adversaries are closing the gap on our information advantage (U.S. Navy 2016). Physical battlespace awareness, or superior knowledge of the operating environment, is an important pillar in maintaining an overall information advantage. As early as Sun Tzu, who identified the knowledge of the "seasons" as a key factor for success against an enemy, strategists have known that awareness of the physical environment is vital to acquiring and maintaining a competitive advantage in warfare (Sun-Tzu 1964).

Naval Oceanography's recent Information Warfare Strategy echoes this theme with the goal of delivering "secure, accurate, timely and precise environmental information in support of Navy Information Warfare" (Gallaudet 2017). Similarly, in 2017 the Oceanographer of the Navy, under the direction of the Chief of Naval Operations, established the Navy's Task Force Ocean. Seeking to maintain our Navy's key advantage in ocean science and Under Sea Warfare (USW), the task force's mission is to assess the state of the Navy's ability to collect, process, and use ocean data (U.S. Navy 2017). The scope of Task Force Ocean extends beyond just the undersea domain. The ability to dominate the ocean battlespace includes the unrivaled ability to predict the physical battlespace from the seafloor to the stars, including the weather. Thus, today's military commanders need the most accurate weather information in order to make better decisions faster than our opponents, especially in situations with high degrees of uncertainty. Naval Oceanography has made great progress in providing critical environmental information to commanders, but progress is still needed in short-term

probabilistic forecasting in regions of limited data availability. This thesis will explore one such challenge in short-term forecasting: ceiling height and visibility.

B. MOTIVATION

Short-term forecasting of sensible aviation weather parameters, such as ceiling height and visibility, continues to prove difficult for the operational forecaster despite significant advances in numerical weather prediction (NWP) and statistical post-processing (Bankert et al. 2004). Ceiling height is defined as the height above ground level (AGL) of the lowest layer of cloud that covers more than half the sky (AMS 2012). Visibility is defined as the greatest horizontal distance at which it is possible to identify an object with an unaided eye (AMS 2012). Critical for aviation safety and military operational planning, accurate and useful short-term forecasts of ceiling height and visibility are necessary to make high-stakes decisions and plans. In order to better impact decision-making processes, forecasts for ceiling height and visibility should be probabilistic in order to better communicate the inherent uncertainty in weather forecasting and allow the decision-maker to explicitly understand all possible outcomes (NRC 2006). Most current operational techniques fail to capture and communicate the inherent uncertainty in ceiling and visibility forecasting. The highly sensitive and chaotic nature of ceiling height and visibility often renders deterministic NWP guidance useless due to limited model spatial and temporal resolution, time latency of availability, and poor communication of uncertainty. Moreover, contemporary statistical techniques require unrealistically large training data sets and fail to communicate useful probabilistic information.

Short-term forecasts of ceiling height and visibility are important because the ability to predict such microscale atmospheric phenomena at short lead times impacts a wide variety of aviation and other military operations. According to the Federal Aviation Administration (FAA), from 1994 to 2003, 21.3% of aviation mishaps were weather-related and one fifth were due to ceiling height and visibility (2017). Moreover, 19% of Class A Navy and Marine Corps aviation mishaps from 1990 to 1998 were weather related, with half related to “visibility related weather elements” (Cantu 2001). Clearly,

despite advances in weather forecasting, poor early warning of low ceilings and visibility continue to add risk and challenges to aviation and military operations. In addition to safety and risk mitigation, accurate predictions of ceiling height and visibility impact military planning and operations. From operating unmanned aerial vehicles (UAVs) and launching F-18 fighter jets off of an aircraft carrier to determining visual weapons ranges, knowledge of the physical battlespace requires the ability to accurately predict the height of the cloud deck and visibility range (JCS 2018). For example, operations at the Naval Strike and Air Warfare Center from May through June 2005, were impacted by weather 24% of the time, with 50% of those (or 12% of all) due to ceiling and visibility (Butler 2005). Furthermore, military forecasters are often placed in forecasting situations with limited regional expertise, observational history, and sensing capabilities. These factors introduce extra challenges in providing accurate and timely forecasts.

The use of probabilistic forecasting can help enhance the ability to issue accurate short-term forecasts because it explicitly communicates uncertainty and mitigates issues of chaos. Conceptually, numerical weather forecasting is an attempt to simulate realistic outcomes of the atmosphere. Based on simplifications of the Navier-Stokes questions and required numerical simplifications and parametrizations, numerical weather models traditionally generate deterministic results that aim to represent the evolution of the real atmosphere. Deterministic, single-valued forecasts, while offering a specific possible outcome, do not communicate the underlying stochastic and chaotic nature of the atmosphere. Lorenz (1963 and 1993) described the inherent sensitivity to initial and boundary conditions and nonlinear growth of error in atmospheric models. Termed Chaos Theory, our inability to accurately determine the initial and boundary conditions of the atmosphere and the nonlinearity of the dynamics limits long-term prediction of the atmosphere. This same atmospheric chaos impacts short-term predictions as well because smaller-scale phenomena entail shorter time periods of predictability due to enhanced turbulence (chaotic dynamics) (Li 2013). At small scales, the intractability of chaotic dynamics leads to necessarily treating some variables as stochastic or random variables. Moreover, because we do not know exactly which (and to what degree) weather variables determine ceiling height and visibility at any given time, we cannot accurately sense the

environment to a high enough spatial and temporal scale to determine the exact initial and boundary conditions. As a result, there may be value in treating ceiling height and visibility themselves as stochastic variables and generating probabilistic versus deterministic forecasts.

Probabilistic forecasts can explicitly communicate this underlying uncertainty of how ceiling height and visibility conditions evolve. Decision-makers need the most accurate and up-to-date weather information in order to make the best decisions. In fact, LeClerc and Joslyn (2015) found that including uncertainty information in weather forecasts led to better decision-making when choosing or not choosing to salt roads during possibly freezing temperatures. For forecasts of ceiling height and visibility, offering probabilistic forecasts of likely ranges or categories allows the consumer to understand the full spectrum of possible outcomes and the associated uncertainty of the forecast based on computational and physical limitations. A probabilistic forecast communicating the most likely outcome, out of a range of possibilities given the current forecasted conditions, would clearly add value to the prediction and allow a commander planning a mission to weigh costs and benefits.

C. SCOPE OF RESEARCH

The goal of this thesis is to demonstrate a prototype system of statistical NWP post-processing for short-term probabilistic ceiling height and visibility forecasting using a Bayesian estimation scheme and Markov Chain Monte Carlo sampling methods. Adapted from Wendt (2017), the machine-learning algorithm constructs realistic distributions of ceiling height and visibility forecasts based on small data sets of North American Mesoscale (NAM) model output near Monterey, CA Airport (KMRY) and corresponding observations from June through August of 2017. The resultant posterior predictive distributions (PPDs) provide the forecaster and decision-maker with statistically rigorous information about the seemingly stochastic nature of the phenomena and the associated uncertainty. The ultimate goal of this research is to incrementally improve the ability to generate accurate short-term probabilistic weather forecasts for aviation safety and military battlespace awareness and decision-making. This study's

hypothesis is that Bayesian post-processed NWP data will reduce error over raw model data and provide useful probabilistic information.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

A. CURRENT FORECASTING TECHNIQUES

Modern techniques used in short-term forecasts of ceiling height and visibility include nowcasting, high-density observation-based statistical forecasting, and NWP post-processing. As a formal technique, nowcasting seeks to predict the state of atmosphere into the near future (less than 6 hours) by extrapolating the current weather based on knowledge of physical principles and observations, including remote sensing products (Mass 2011). Nowcasting can be accurate in certain situations, such as forecasting the advance of a squall line or mesoscale convective complex. In fact, Dixon (2004) demonstrated the feasibility of a system for nowcasting visibility using observations and RADAR reflectivity. Using an empirical relationship between reflectivity and visibility, the system extrapolates current reflectivity to determine visibility in snow conditions. Modern nowcasting methods, such as the NCAR Auto-Nowcast System, blend observations, remote sensing data and mesoscale models in order to generate short-term forecasts (Mueller et al. 2003). However, nowcasting is limited by chaos in the atmosphere, our incomplete knowledge of the atmosphere at the turbulent scale, the individual ability of the forecaster, and the limited availability of observations and remote sensing tools and products. Moreover, military forecasters in remote regions will always face challenges in regional expertise as well as data, sensor, and mesoscale model availability.

Next, high density observations-based systems use a robust network of surface and upper-air weather observations as predictors in a statistical forecasting system. Vislocky and Fritsch (1997) demonstrated a highly successful technique using observations as both predictors and predictands and least squares multiple linear regression to generate short-term probabilistic forecasts that outperformed model output statistics (MOS) out to six hours. Application of this technique at San Francisco International Airport yielded useful results and a 32% reduction in mean square error compared to MOS (Hilliker and Fritsch 1999). This technique can be highly accurate, but

it requires availability of a dense network of observations in order to add value beyond MOS (Leyton and Fritsch 2004). Furthermore, this technique ignores errors in the predictors and does not necessarily eliminate bias (Marzban et al. 2006). Over the continental United States with large training sets of observational history, this technique is realistic but not in remote regions across the globe.

Statistical post-processing seeks to link raw NWP model output with observations to find linear or nonlinear relationships. Ryerson (2012) analyzed mesoscale visibility forecasts using uncalibrated ensembles and identified errors associated with initial conditions, parametrizations, and errors in deriving visibility from forecasted variables. He concluded that "under most conditions," post-processing is required to generate useful forecasts. The canonical example of statistical post-processing is Model Output Statistics (MOS). It is an attempt to reduce bias, limit error variance and reduce mean square error (Marzban et al. 2006). Developed by the National Weather Service (NWS) in 1972, MOS is a statistical post-processing technique that uses a list of predictors, both observations and model output, that are chosen based on correlations to the predictands (in this context ceiling height and visibility) in a multiple linear regression scheme (Glahn and Lowry 1972). The regression scheme determines appropriate regression coefficients and develops the linear regression equation for forecasting. The MOS forecast output for ceiling height and visibility is a categorical forecast that attempts to adjust for model bias and local affects (Bocchieri and Glahn 1972). Table 1 lists the forecast categories and related aviation flight categories. The aviation flight categories are low instrument flight rules (LIFR), instrument flight rules (IFR), marginal visual flight rules (MVFR), and visual flight rules (VFR) (NWS 2010).

Table 1. Ceiling Height and Visibility Categories and Associated Aviation Flight Categories. Adapted from NOAA, http://www.nws.noaa.gov/mdl/synop/namcard.php

Ceiling Height			Visibility		
Category	Height (ft)	Aviation Flight Category	Category	Range (miles)	Aviation Flight Category
1	<200	LIFR	1	<.5	LIFR
2	200-500	LIFR	2	.5-1	LIFR
3	500-1000	IFR	3	1-2	IFR
4	1000-2000	MVFR	4	2-3	IFR
5	2000-3000	MVFR	5	3-5	MVFR
6	3000-6500	VFR	6	5-6	VFR
7	6500-12000	VFR	7	>6	VFR
8	>12000	VFR			

Similarly, the Localized Aviation Model Output Statistics Program (LAMP) downscales the MOS process and products for specific airfields, provides hourly updates to the MOS forecast production cycle, and also provides probabilities of different ceiling height and visibility categories (Rudack and Ghirardelli 2010). This method is currently the standard forecast product at NWS Weather Forecast Offices (WFO) for Terminal Aerodrome Forecasts (TAFs) at U.S. airfields. While statistically robust and tailorable, the MOS and LAMP forecast products do not communicate the associated uncertainty, such as error or confidence intervals. Moreover, development of the regression questions requires large data sets of model output and observations, both of which are unrealistic in military operation scenarios. Finally, this system is not available at worldwide airfields, let alone in remote regions.

Advanced machine learning techniques include using artificial neural networks to forecast ceiling and visibility that allow for nonlinear relationships and use continuous feedback loops to update forecasts (Marzban et al. 2006). Similarly, data-mining techniques at single-stations have shown promise at accurate short-term forecasts (Bankert et al. 2004). Fuzzy logic systems, advanced by Hansen (2007), use the idea of partial truth, as opposed to Boolean logic, to collect similar analogs to make ceiling and

visibility predictions. Other advanced statistical post-processing techniques include Bayesian Model Averaging (BMA). This technique, as demonstrated by Chmielecki and Raftery (2011), calibrates forecast ensembles and produces predictive distributions by combining mixed conditional probability distributions of predictors. All of these techniques show great promise but still contend with the severe environmental sensitivity in ceiling height and visibility processes and the dearth of requisite historical model data and observational history for model training.

B. BAYESIAN INFERENCE IN WEATHER FORECASTING

This thesis will apply a Bayesian Estimation (BE) scheme and Monte Carlo Markov Chain (MCMC) sampling methods scheme adapted from LCDR Travis Wendt's doctoral research in order to generate short-term probabilistic forecasts of ceiling height and visibility. In simple terms, Wendt's model formalizes forecasting heuristics by discovering the most significant predictors that influence the predictands. Using a limited training period, Bayesian thinking, and MCMC to complete the inference, the model uses stochastic multivariate multiple linear regression to learn the appropriate distributions of regression coefficients that most efficiently explain the relationships between the training data predictors and predictands (Wendt 2017). Wendt's research application focused on forecasting weather parameters for the North American collegiate weather forecasting competition called The Weather Challenge. He successfully used predictor variables derived from the NCEP Short Range Ensemble Forecast (SREF) ensemble means to forecast PPDs of 24-hour maximum temperature, minimum temperature and maximum wind speed (Wendt 2017). Using a limited training data set of just one year, Wendt's model matched or exceeded forecast reliability (forecasted probability of occurrence is close to actual occurrence) of the unprocessed SREF output (Wendt 2017). His research demonstrated the ability to use short training periods to generate useful and competitive probabilistic forecasts. Details about the theory and mechanics of this model will be discussed in the next chapter.

III. DATA/METHODOLOGY

A. OVERVIEW

In general, the ceiling height and visibility nowcast system of this research post-processes raw mesoscale NWP model output by finding the appropriate distributions of regression coefficients in a stochastic multivariate multiple linear regression scheme. The system can produce hourly probabilistic forecasts of ceiling height and visibility by finding the statistical relationship between past observations and predictors. What differentiates this system's method and standard multiple linear regression schemes, such as MOS, is that the model is finding distributions of regression coefficients, instead of deterministically valued coefficients, and therefore generates predictive posterior distributions (PPD). Specifically, Bayesian Estimation is used to invert the probability statement to find the probability of the model parameters given training data. Markov Chain Monte Carlo methods are used to complete the inference between multiple predictor variables and the multivariate predictands. Multiple combinations of predictor variables are used and compared against raw NWP model predictions. K-means clustering is used to logically partition the training and test data in order to minimize variance. Although this research attempts to post-process NWP output at a single grid point near Monterey Regional Airport (KMRY), the method can be applied to multiple grid points to develop a gridded forecast. Errors are then compared between raw NWP forecasts and the mean and median of the PPDs. The forecast trial with the set of predictors that produce the smallest error is scored for skill over the raw NWP forecast. Finally, although not scored, the probabilistic capability of the system is demonstrated by considering the PPD credible intervals and categorical probabilities.

B. DATA SET

1. Automated Surface Observing System (ASOS) History

To develop the training data and scoring observational history, ASOS observations are used from KMRY from June to August 2017. The station elevation is

257 ft above mean sea level (MSL). Hourly surface observations are retrieved from the National Centers for Environmental Information (https://www.ncei.noaa.gov). The data is downloaded in text format and decoded using MATLAB. The date time group (DTG), ceiling (CLG), and visibility (VIS) variables are saved and the remaining data is discarded. The ceiling data is encoded in hundreds of feet up to 12,000 ft. For example, a CLG observation of "12" is decoded at "1200 ft" above ground level (AGL). Visibility data is encoded in statute miles up to 10 miles. For example, a visibility observation of "10.0" is decoded at 10 statute miles. The ceiling and visibility observations are the predictands (the variables to be predicted) in the model. A total of 3369 hourly observations were pulled from June 1, 2017 0100Z to August, 31, 2017 2300Z.

2. NWP Model Output

The predictor variables were extracted from the 12 km resolution North American Mesoscale (NAM) model at the grid point located at 36.65N 121.74W, about 7.36 miles north east of Monterey Regional Airport. This grid point was chosen because it is the nearest geographically similar grid point. Figure 1 portrays the location of the grid point at 36.65N 121.74W and the airport. Table 2 lists the variables obtained from the NAM model for investigation as predictors. A total of 2142 hours of NAM model output was extracted from Gridded Binary files (GRIB). Hourly forecasts extended out several hours in time, or taus, of 0 to 11. Not all hours of the time period were extracted due to various errors in the model output.

Figure 1. Map of Monterey Bay.

Table 2. Extracted NWP Variables

Level (mb)	Variables (units)
surface	mean sea level pressure (mb); temperature (F); dew point (F); wind (kts); wind gusts (kts); height of planetary boundary layer (m); precipitation (in); surface pressure (mb); visibility (km); u wind component (m/s); v wind component (m/s); cloud cover (%); cloud base height (m)
500, 700, 850, 875, 900, 925, 950, 975, 1000	height (m); temperature (C); relative humidity (%); dew point (C); wind speed (m/s); wind direction (degrees); u wind component (m/s); v wind component (m/s); vertical velocity (m/s)

C. DATA QUALITY CONTROL AND MERGING

Initial quality control of the data was first accomplished in MATLAB. ASOS observations without ceiling height or visibility readings were discarded. After this initial data cleaning, 2950 observations were left. Next, the NAM predictor data was quality controlled. Hours missing model data were discarded which left 1972 hours of data. Finally, the observational data and model data were joined based on date time group. This produced a full merged data set of 1379 hours of observation and model data. The first 1183 hours were used at training data and the last 196 hours were used as test (forecast) data.

D. CORRELATIONS AND CLUSTERING

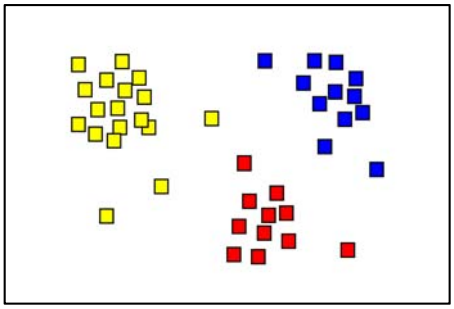
1. Correlations

In order to choose the most influential predictors, linear correlations were run on each predictand-predictor relationship. First, the NAM cloud base height predictor was adjusted from above MSL to AGL by subtracting 257 ft (KMRY ASOS station height) from the cloud base height forecast. This ensures that the heights are consistent since the ASOS ceiling height is reported as height AGL. Next, Pearson correlation coefficients were used to characterize the linear relationships. Correlations were accomplished on the full data set as well as the data set conditioned on ceiling height observations of less than 6000 ft. This was necessary because the ceiling height observations were artificially capped at 12200 ft by the ASOS observations. This means that any ceiling above 12000 ft was coded as 12200 ft. Similarly, visibility observations were capped at 10 miles. This drastically limits the ability to find linear relationships between predictors and predictands for high ceilings and extended visibility ranges. These issues will be discussed more in the results section.

2. Clustering

K-means clustering was used to cluster both the training and test data. In general, clustering is used to group similar data based on specified rules. The goal of clustering is to find the groupings that minimize variance and lead to error reduction in predictive analytics (Pedregosa et al. 2011). Clustering is used in pattern recognition, data mining, and supervised and unsupervised learning to find the hidden patterns in data (Pedregosa et al. 2011). In 2 and 3 dimensions, the human eye can visually group most data, given there are clusters. In Figure 2, it is readily apparent there are 3 clusters of data.

Figure 2. Simple Clustering in 2 Dimensions. Source: “Cluster Analysis, Wikipedia, https://en.wikipedia.org/wiki/Cluster_analysis#/media/File:Cluster-2.svg



Another simple example in 2 dimensions is identifying patterns between low and high ceiling heights. Figure 3 portrays the joint plot of ceiling height in feet (x-axis) and surface temperature in degrees F (y-axis) for data from this study. The associated histograms and kernel density estimation are plotted along the opposite respective axes (ceiling height on top and temperature on far right). The Pearson correlation coefficient is r=0.45. It is apparent that there are two clear clusters of ceiling height data: low and high. The logical break is around 5000–6000 ft. This bimodal distribution affects the linear regression between the two variables. The problem with ceiling height observations capped at 12200 ft is apparent here where there is a vertical line of data on the right-hand side of the plot. This also severely skews the regression. To easily adjust, we can manually cluster the data by conditioning the regression on ceilings less than 6000 ft. This eliminates the bimodal distribution and changes the regression. Figure 4 portrays the conditional regression with r=.019. By manually clustering 2-D data based on a single dimension (ceiling height), the regression is dramatically adjusted. The interpretation is that surface temperature is not a good predictor of ceiling heights of less than 6000 ft, based on available data and assuming a linear relationship.

Figure 3. Joint Plot of Ceiling Height and Surface Temperature

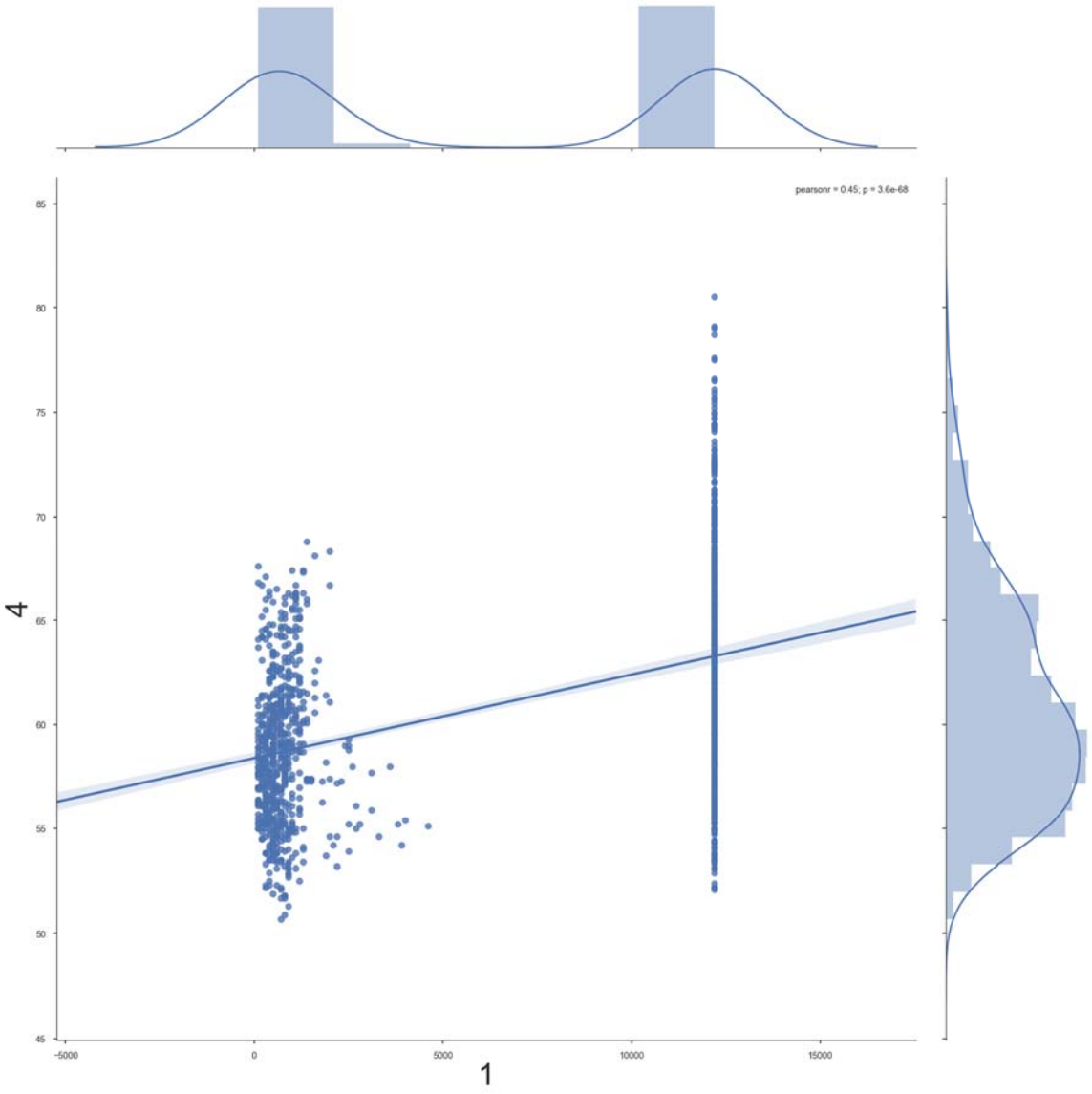
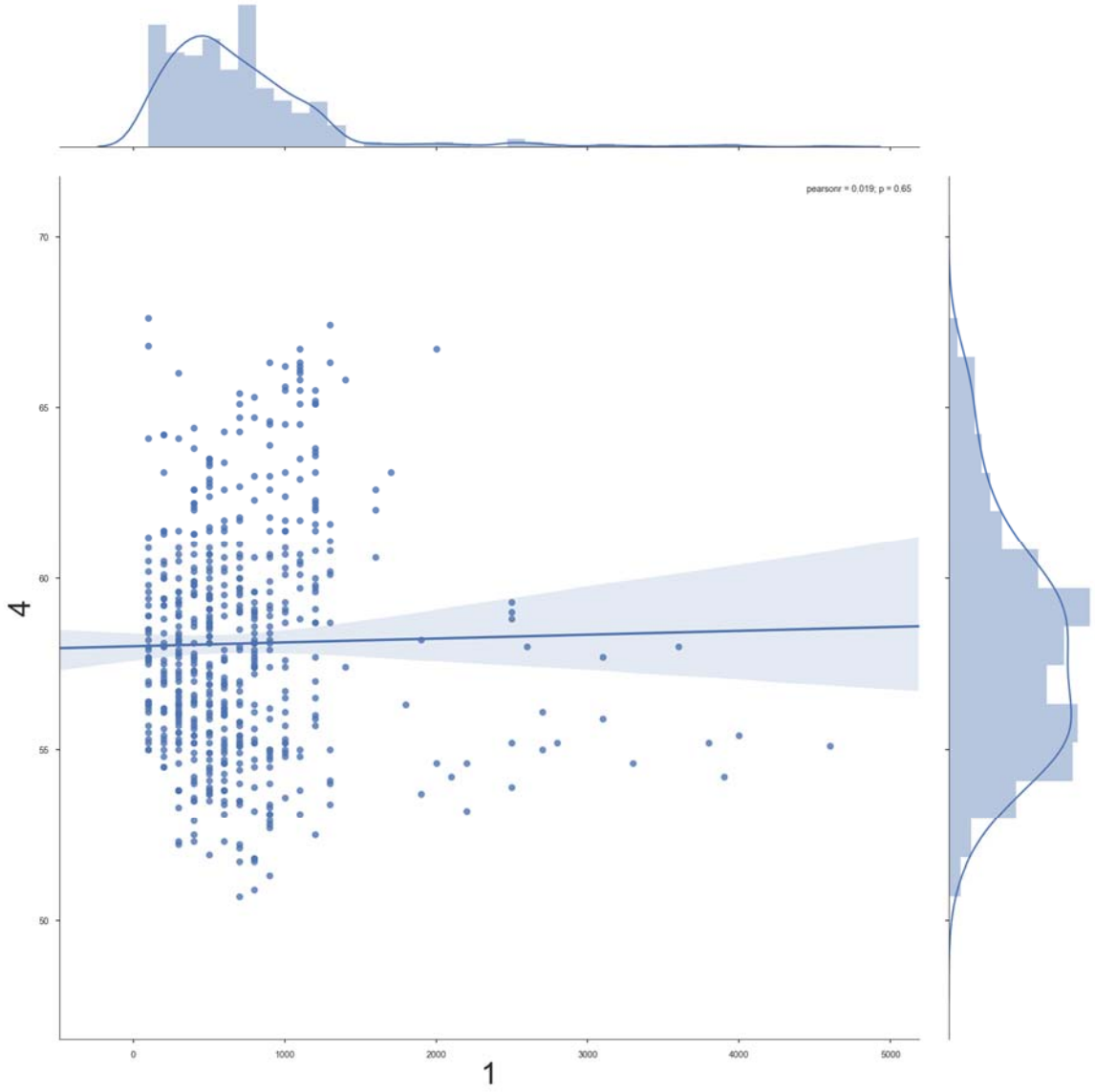


Figure 4. Joint Plot of Ceiling Height Less than 6000 ft and Surface Temperature



In more than 3 dimensions, humans cannot cluster data because we cannot visualize 4 dimensions and higher. To do this, clustering algorithms can accomplish the task using a variety of methods. This thesis utilizes K-means clustering, which minimizes

the variance, or the sum-of-squares, within the cluster from its Euclidean center of mass (Pedregosa et al. 2011). This method requires the user to specify the number of clusters, and the algorithm then finds the most efficient partitioning of the data to minimize the distance between each data point and the clusters' centers of mass. K-means randomly chooses the centers of mass and then iterates repeatedly until convergence where the variance, or inertia, is minimized (Pedregosa et al. 2011).

K-means does not identify the optimal number of clusters itself (Pedregosa et al. 2011). In order to determine the optimal numbers of clusters, this thesis employs silhouette analysis to subjectively choose the number of clusters. Silhouette analysis measures the distances between clusters and assigns the cluster a silhouette score from -1 to 1. A value of 1 means the clusters is clearly distinct and separate from nearby clusters. A value of 0 means the cluster is very nearby a neighboring cluster. Negative values mean data may have been assigned to the wrong cluster (Pedregosa et al. 2011). High scores close to 1 with low variability are desirable. The score is a representation of how different numbers of clusters successfully group that data. The shape of the silhouette communicates relative sizes or density of the clusters (Pedregosa et al. 2011). The user identifies the range of clusters for analysis. In addition, for each silhouette analysis, this study compares the silhouette score distribution for 100 samples. This allows the user to understand the uncertainty in the analysis by repeating it many times and considering the variability of the mean silhouette scores for each k, or number of clusters (Zoufaly 2017).

For example, Figure 5 displays the mean score distribution for k=2-8 clusters using 2 predictors and 2 predictands for a total of 4 dimensions. The distributions consist of 57 and 28 subsamples from a sample of 100 cluster analyses. Based on subjective visual score analysis alone, k=3 or k=4 appears to be the most optimal number of clusters because they each have low mean variability, especially at 57 subsamples. Figure 6 portrays the silhouette scores for 10 trials (left image) and the 2-D plot (right image) of the clustered data for the 2 predictand dimensions (ceiling height and visibility observations). The x-axis of the left image measures the silhouette coefficient or score from -1 to 1. The red vertical line is the average score for the 10 trials. The y-axis of the

left image identifies the cluster number. Looking at N=3 clusters, all clusters are above the average score and they have similar density (shape). Clusters 1 and 2 (do) have some data that may be assigned to the wrong cluster as shown by data less than 0. The 2-D plot shows that clusters 1 and 3 are near each other but are distinguished by the other 2 dimensions. Looking at N=4 clusters, all clusters are above the average score, but cluster 3 has a wildly different shape and lower density than the other 3. Cluster 2 has a significant number of data possibly assigned to the wrong cluster. Moreover, three of the clusters appear to be near each other, at least in two dimensions, based on the 2-D plot. Based on this subjective analysis, the plot for k=3 clusters appears to be the optimal number of clusters to use for partitioning the data.

Figure 5. Violin Plot of Silhouette Score Distribution

Figure 6. Silhouette Analysis, N=3 Clusters

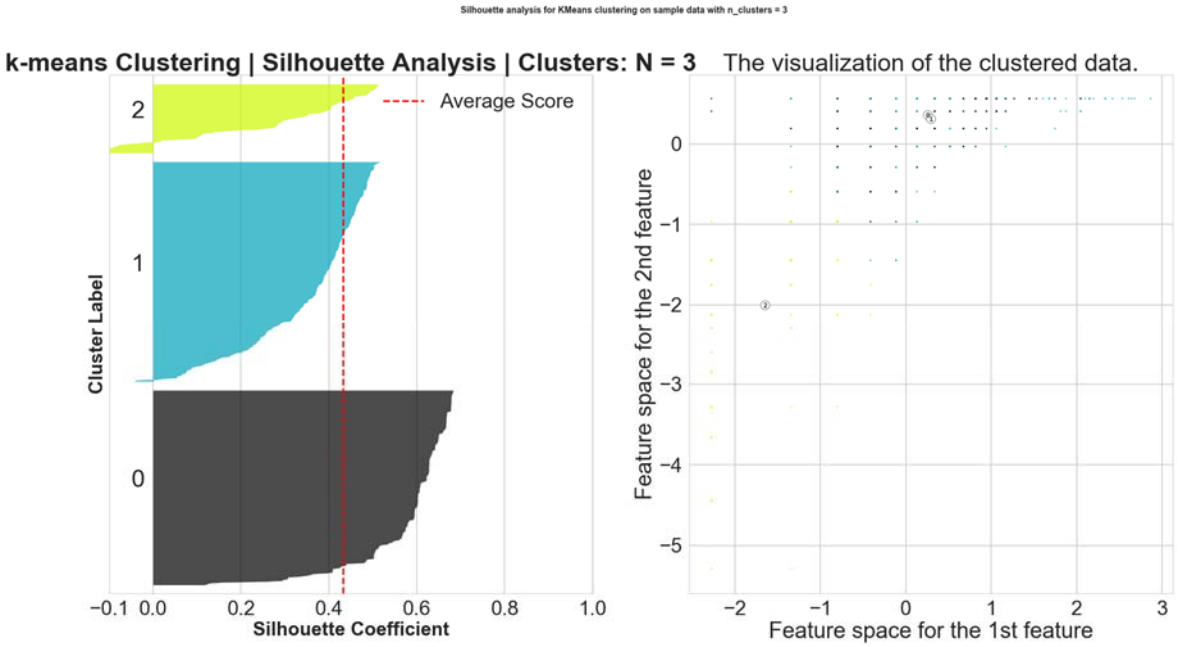
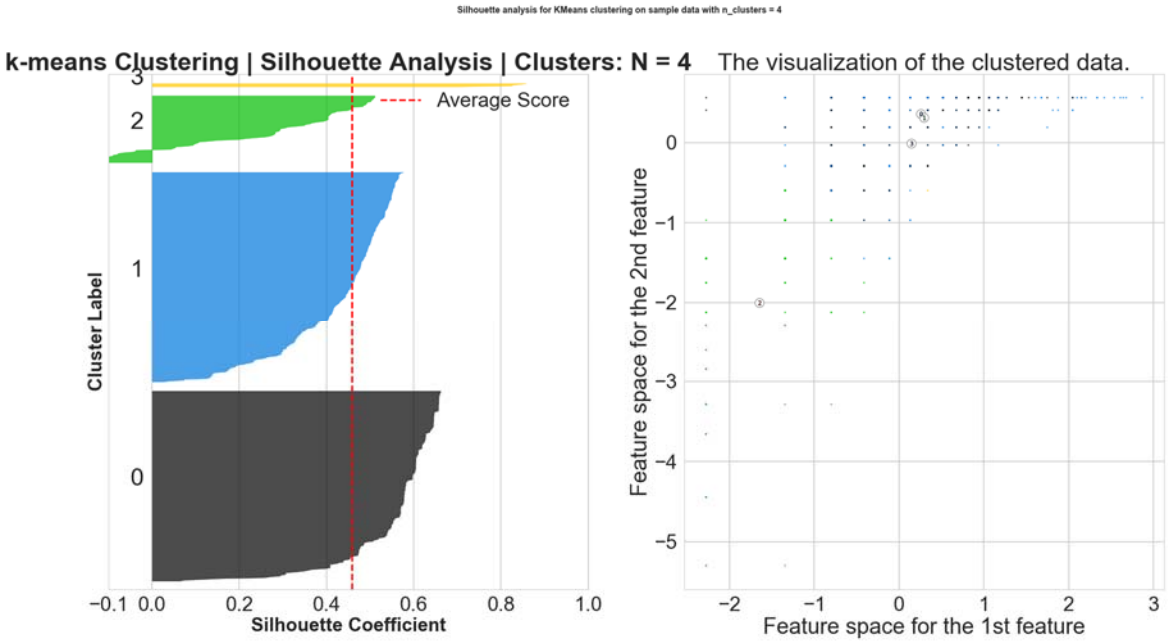


Figure 7. Silhouette Analysis, N=4 Clusters



For this study, K-means clustering was used to cluster up to 8 dimensions: 2 dimensions of predictands and 6 dimensions of predictors. The training set was clustered on all 8 dimensions. In order to ensure the training data, which includes the observational history, does not bleed into the test data, the 2 predictand dimensions in the test data are replaced by the respective raw model predictions in the test set clustering. This operation introduces some error because the choice of raw NWP forecasts to replace the predictand dimensions is an imperfect inference. In this way, the use of k-means to organize the training and test data into optimal clusters can be considered a limited case of supervised machine learning (Pedregosa et al. 2011).

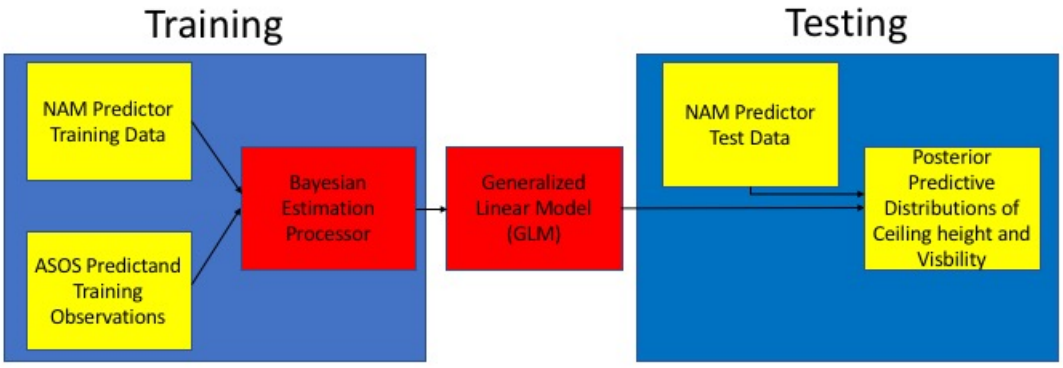
E. BAYESIAN POSTPROCESSING METHOD

1. Concept

This thesis applies Bayesian Estimation and Monte Carlo Markov Chain sampling methods adapted from Wendt (2017). This method explicitly applies Bayes' Rule in order to determine information about the model parameters given the training data. Predictor

and predictand training data is used to develop generalized linear model (GLM) parameters via Bayesian Estimation, which is then used to forecast ceiling height and visibility given test predictor data in the test or forecast step. The forecast is communicated via a posterior predictive distribution (PPD). The mean or median of the distribution or the calculated cumulative probability for certain conditions can be used to generate specific forecasts. Conceptually, the model determines the sensitivity of the predictands to the chosen predictors, and then uses that knowledge to make predictions probabilistically. In this way, the model quantifies forecasting thumb-rules, or heuristics. Figure 8 summarizes the process visually.

Figure 8. Bayesian Post-Processing Model Schematic



2. Bayesian Estimation and MCMC Method

Equation 1 details Bayes' Rule, which inverts the conditional probability statement. The rule states that the probability of model parameters θ, given training data Y, is equal to the probability of the data Y, given the model parameters θ multiplied by the probability of the model parameters θ, and divided by the probability of the data Y. $p(\theta|Y)$ is known as the posterior probability or belief in Bayesian terms. $p(Y|\theta)$ is known as the likelihood function. $p(\theta)$ is known as the prior probability. $p(Y)$ is the marginal probability, or evidence, of Y and serves as a normalizing factor (Wilks 2011).

$$p(\theta \mid Y) = \frac{p(Y \mid \theta)\, p(\theta)}{p(Y)} \qquad \text{Equation (1)}$$

The power of this model is that Bayesian thinking is superior to frequentist thinking for scientific tests and predictions. Traditional frequentist statistics typically make predictions in terms of $p(Y|\theta)$. In words, the prediction is the probability of forecast data Y, given the parameters θ of a normal likelihood distribution $N\sim(\mu,\sigma^2)$ derived from a set of sample data. The misuse of frequentist statistics in the task of scientific predictions is well-known and documented by data scientists (VanderPlas 2013). The simplest way to understand the difference between Bayesian statistics and frequentist statistics is to examine the differences in the way each communicates uncertainty. Uncertainty is typically communicated via credible intervals in Bayesian statistics and confidence intervals in frequentist statistics. Both intervals are calculated in the same way by adding and subtracting a multiple (1 for 68.27%, 2 for 95.45% and 3 for 99.72%) of the standard deviation from the mean (VanderPlas 2013). The difference is in the interpretation and mathematical meaning. Bayesian statistics treats the bounds of the intervals as fixed, and the sample parameters μ and σ as random variables (VanderPlas 2013). Thus, when communicating the 95% credible interval, for example, it is accurate to state that there is a 95% probability that the sample parameter μ falls within the interval bounds. This is sound reasoning because the Bayes' Theorem computes the probability of model parameters given training data. The intervals are constructed by the training data while the prediction is a probabilistic statement about the belief that a trial's parameters will fall within the expected range. On the other hand, frequentist statistics treats the sample parameters μ and σ, derived from the sample statistics, as fixed and treats the bounds of the interval as random (VanderPlas 2013). Thus, stating that there is a 95% chance of a trial's parameters falling within the 95% confidence interval is not correct reasoning. Confidence intervals refer to the probability that if the experiment is repeated 100 times, 95 of the computed intervals will contain the true mean and standard

deviation. It is also important to note that when calculating credible intervals, if a normal likelihood function and non-informative prior $p(\theta) = 1$ are used, then the credible interval and confidence interval will be numerically exactly the same (VanderPlas 2013). So, while credible intervals and confidence intervals are computed exactly the same and are often numerically similar, they have drastically different meanings (VanderPlas 2013). When making scientific predictions, as in weather forecasting, it is more accurate to use Bayesian statistics and credible intervals.

In this model, θ is the set of regression coefficients corresponding to each intercept, predictor weight, variance and covariance. This model forecasts for 2 predictands, ceiling height and visibility, using combinations of 2–6 predictors. So, for a trial with 2 predictors, there will be 9 regression coefficients. The 9 regression coefficients consist of the intercepts for each predictand (θ_0 and θ_1), a coefficient for each predictor (θ_2-θ_5), the variances of the 2 predictands respectively (θ_6-θ_7), and the predictands' covariance θ_8. During the training step, *Y* is the observed ceiling height and visibility observations. During the test step, *Y* is the prediction for ceiling height and visibility. In this way, this application of the Bayesian post-processing model seeks to make PPDs of *Y* test data inferred from a GLM based on past training data.

The complexity of Bayesian estimation lies in the method to complete the inference. Looking back at Equation 1, $p(Y)$ is not always known, and it must be removed to complete the task.

$$p(\theta \mid Y) = \frac{p(Y \mid \theta)\, p(\theta)}{p(Y)} \tag{1}$$

So, knowing that $p(Y) = \int p(Y|\theta)p(\theta)d\theta$, Equation 1 can be rewritten as:

$$p(\theta \mid Y) = \frac{p(Y \mid \theta)\, p(\theta)}{\int p(Y \mid \theta)\, p(\theta) d\theta} \tag{2}$$

Equation 2 is the inference needed to successfully calculate the probability of parameters θ, given training data Y from the likelihood function $p(Y|\theta)$ and the prior $p(\theta)$. For this study, no assumption is made about the prior so $p(\theta)=1$ and $p(Y|\theta)$ is assumed to be a normal likelihood function, or the assumption about how the data is distributed. In order to complete the inference of equation 2, MCMC methods are used to estimate the distribution. MCMC method are necessary due to the multivariate predictands and multiple predictors that make the inference computationally intensive (Gelman 2013). Specifically, this study utilizes the Metropolis algorithm, which creates the Markov Chain using random jumps or walks (Gelman 2013). By taking random jumps, the process makes no assumptions about the path of the Markov Chain. For each regression parameter θ, the algorithm computes $p(\theta_n|Y)$ at the current location in sample space θ_n and then randomly jumps to a new location and computes $p(\theta_{n+1}|Y)$ at θ_{n+1}. Equation 3 is the ratio of the 2 jumps. Notice, $p(Y)$ cancels, which eliminates the problem of the denominator in equation 1.

$$r = \frac{p(\theta_{n+1}|Y)}{p(\theta_n|Y)} = \frac{p(Y|\theta_{n+1})p(\theta_{n+1})p(Y)}{p(Y|\theta_n)p(\theta_n)p(Y)} \tag{3}$$

When the ratio r is greater than a randomly drawn number from the interval (0,1), the algorithm accepts the jump to the location of phase space of θ_{n+1} (Wendt, 2017). If not, the jump is rejected and the algorithm retains the location of θ_n as the next step in the chain. Acceptance of the jump is not guaranteed so as to ensure the chain visits the locations of sample space with higher posterior probability more often than regions of low posterior probability and maintains balance (Wendt 2017). The accepted jumps build the posterior sampling distribution. Remembering that the parameters θ represent the regression coefficients in the GLM, the chain of states of θ in the sample space is the Markov Chain representing the distribution of respective regression coefficients. The algorithm generates a Markov Chain for each regression parameter θ in the GLM. Each chain converges when the sampling distribution reaches stationary behavior (Wendt 2017

and Gelman 2013). In this application, the model computes $1x10^6$ proposals for each chain. After the samples converges, or after $5x10^5$ samples and half the chain, the model randomly draws $1x10^4$ samples of the chain to build the discrete posterior distribution. Kernel Density Estimation (KDE) is used to estimate the continuous probability distribution form of the discrete histogram by fitting a normal distribution to each sample and then linearly summing to produce a continuous distribution (Wendt 2017). Each parameter θ corresponds to regression coefficients called betas (β) in the GLM.

For example, for a model with two predictands Y_1 and Y_2 and two predictors X_1 and X_2, the GLM is described in equation set 4.

$$\begin{aligned}
&\vec{Y}(X \mid \beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7,\beta_8) \sim N(\vec{\mu},\Sigma) \\
&\mu_1 = \beta_0 + \beta_2 X_1 + \beta_3 X_2 \\
&\mu_2 = \beta_1 + \beta_3 X_1 + \beta_5 X_2 \\
&\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}
\end{aligned} \tag{4}$$

Here, $\vec{Y}$ refers to a vector containing the 2 predictands, ceiling height and visibility, which is a function of the predictors X given parameters $\beta_{0\text{-}8}$ ($\theta_{0\text{-}8}$). The GLM of $\vec{Y}$ is approximated by a normal distribution with parameters $\vec{\mu}$ and Σ. The vector $\vec{\mu}$ contains the parameters μ_1 and μ_2, which are defined as linear summations of the regression parameters (betas) and the predictors. The parameter Σ is the covariance matrix containing the variances σ_1^2 and σ_2^2 of the predictands on the diagonal plus the product of covariance and correlation coefficient of the predictands: $\rho_{12}\sigma_1\sigma_2$. Note, the covariance term is redundant in the off-diagonals. Parameters $\beta_{6\text{-}8}$ are the variance parameter estimates of the covariance matrix Σ. Thus, they are not literally the values of the variances and covariance. Cholesky decomposition must be performed to retrieve the actual values (Wendt 2017). In this context, it is not necessary to perform the

transformation. Unlike simple linear regression, this GLM includes the variance and covariance of the predictors in the regression.

Finally, in the forecast or test step, test data predictors are input into the GLM. The model is trained with log-transformed and z-scored (standardized) data, and, therefore, the forecast includes exponentiation and un-standardization of the posterior. The log-transformation is necessary to handle the relatively non-normal distributions for the predictors by forcing them to be more normal, to ensure forecasts are positively-valued (no negative ceiling heights and visibility forecasts), and to produce asymmetric PPDs that capture the true and intrinsic skewness of the weather variables (Wendt 2017). The z-score transformation (standardization) is necessary because the multiple predictors and predictands are in different units and scales (Wendt 2017 and Pedregosa *et al.* 2011). The transformation standardizes each data point by subtracting the mean and dividing by the standard deviation, which produces unit-less z-scores. Finally, the forecasts are transformed back into their respective original units in the test step. Lastly, the mean and median of the posterior distributions are calculated for scoring purposes against the raw NAM forecast. Credible intervals are then calculated to communicate uncertainty.

In summary, test predictors are inputted into the GLM trained on past relationships for the two predictands. The output is a predictive posterior probability distribution. Through Bayesian reasoning, the interpretation of the posterior distribution and credible intervals correctly reflect the probability, or belief, of the ceiling height and visibility predictions.

F. SCORING

For each run, the mean and the median of each prediction is calculated. Next, mean error (ME), mean absolute error (MAE), root mean square error (RMSE) and skill score (SS) are calculated for both the mean and median forecast sets. Equations 5–8 list the equations for ME, MAE, RMSE and SS.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(f_i - t_i) = \frac{1}{n}\sum_{i=1}^{n}(e_i) \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\ f_i - t_i\ | = \frac{1}{n}\sum_{i=1}^{n}|\ e_i\ | \quad (6)$$

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}(f_i - t_i)^2\right)^{1/2} = \left(\frac{1}{n}\sum_{i=1}^{n}(e_i)^2\right)^{1/2} \quad (7)$$

$$SS = 1 - \frac{RMSE_{NPS}}{RMSE_{NAM}} \quad (8)$$

In the calculations, n is the number of forecasts, f_i is the mean or median of ith forecast, t_i is the corresponding truth from the observation and e_i is the error. Next, the ME, MAE, and RMSE is calculated for the raw NAM forecasts of ceiling and visibility and compared to the model's post-processed errors. Mean error indicates the model's average simple performance. Mean absolute error calculates the overall average error without distinguishing between under and over-forecasting. Root mean square error is similar to MAE but it penalizes severe individual errors more. Skill score represents the percent improvement (or degradation) in RMSE. All four error calculations are standard scoring metrics in meteorology (Wilks 2011). In addition to being calculated on the full set of forecasts, the errors are also calculated on the subset of forecasts corresponding to the aviation flight categories (table 1) for Run 8a (the best performing trial).

Next, forecast skill for Run 8a is scored against NAM using contingency tables and associated metrics conditioned on the aviation flight categories. An example contingency table is shown in table 3. A contingency table tabulates the number or ratio of correct forecasts compared to observations (Wilks 2011). For example, this study compares the number of times LIFR ceiling height conditions are forecasted by the model

(NAM or NPS) and LIFR do conditions occur (box a). Second, the number of times LIFR conditions are not forecasted (some other condition is forecasted) and they do occur is counted (box c). Third, the number of times LIFR conditions are forecasted and they do not occur is counted (box b). Fourth, the number of number of times LIFR conditions are not forecasted and some other conditions occurs is counted (box d).

Table 3. Contingency Table Example. Adapted from "Verification Measures," http://www.wxonline.info/topics/verif2.html

LIFR		Observed		
		Yes	No	
Forecasted	Yes	a	b	a+b
	No	c	d	c+d
		a+c	b+d	n=a+b+c+d

From this table, standard weather verification metrics are calculated to include: percent correct (PC), hit rate (HR), false alarm ratio (FAR), threat score (TS), bias, and Heidke Skill Score (HSS). Equations 9-14 list the mathematical calculation for the example in table 3.

$$PC = \frac{(a+d)}{n} \tag{9}$$

$$HR = \frac{a}{(a+c)} \tag{10}$$

$$FAR = \frac{b}{(a+b)} \tag{11}$$

$$TS = \frac{a}{(a+b+c)} \tag{12}$$

$$Bias = \frac{(a+b)}{(a+c)} \tag{13}$$

$$HSS = \frac{2(ad-bc)}{[(a+c)(c+d)+(a+b)(b+d)]} \tag{14}$$

The percent correct (PC) calculates the percent of the forecasts that are correct with 1 being perfect (Wilks 2011). The hit rate (HR) is the ratio of observed events that are forecast with 1 being perfect (Wilks 2011). The false alarm rate (FAR) is the rate of yes forecasts that were incorrect with 0 being best. The threat score (TS) combines HR and FAR into one score with 1 being best (Wilks 2011). Bias compares the number of times a condition was forecasted and the number of times it occurred. A bias=1 means the event occurred as many times as it was forecasted. A bias<1 means a condition was forecasted less times than it occurred. A bias>1 means a condition was forecasted more times than it occurred (Wilks 2011). The Heidke Skill Score (HSS) compares the proportion correct to that which would occur from random forecasts that are independent of the observations. An HSS=1 means the forecast is perfect compared to a random forecast, HSS=0 means no skill over random forecasting, and HSS<0 means worse than a random forecast (Wilks 2011). Finally, to better understand the probabilistic capability of the model, posterior predictive distributions from a select forecast from Run 8a is presented and discussed.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RESULTS

A. PREDICTOR SELECTION AND DATA CONDITIONING

1. Correlations for Unconditioned Data

Predictor selection was based on 2 approaches. First, the raw NAM forecasts for cloud base height (index 13) and visibility (index 11) were used as predictors to test this nowcast system's ability to post-process explicit NWP cloud base height and visibility forecasts. It is important to note that cloud base height and ceiling are not the same thing. As defined earlier, ceiling height is the height of the lowest layer of clouds covering at least half the sky (AMS 2012). Cloud base height above mean sea level (MSL) is the height of the lowest layer of clouds covering any portion of the sky (AMS 2012). As discussed previously, the height of the station must be subtracted from the forecast to maintain consistency of height reference. Thus, NAM forecasts of low cloud base may not necessarily generate a ceiling and are modified by the elevation of the station. However, the cloud base height forecast is a good approximation of ceiling height given that a ceiling exists.

Second, predictors were chosen based on correlations to both ceiling height and visibility predictands as well as their physical relevance. The second set of predictors does not include the raw NAM cloud base height and visibility forecasts in order to investigate the ability to make predictions using the relationships between ceiling height and visibility and physical variables other than the explicit forecasts. For this study, the physical variables chosen based on high correlation coefficients and physical relevance to the phenomena include: surface wind (index 6), 950mb relative humidity (index 39), 975mb relative humidity (index 42), 975mb wind speed (index 43), 1000mb relative humidity (index 46), and 1000mb wind speed (index 47). Pearson correlation coefficients for the raw data (unconditioned) are shown in Table 4.

Table 4. Correlation Coefficients for Raw Predictors

Index	Level	Variable Name (units)	Ceiling r	Visibility r
6	surface	Wind speed (knots)	0.331583	0.343562
11	surface	Visibility (km)	0.100516	0.085129
13	surface	Cloud base height (m)	0.461871	0.114438
39	950	Relative humidity (%)	0.354242	0.075211
42	975	Relative humidity (%)	0.522398	0.064485
43	975	Wind speed (m/s)	0.20513	0.327587
46	1000	Relative humidity (%)	0.549819	0.149936
47	1000	Wind speed (m/s)	0.312252	0.329551

Figures 9-12 display the joint plot (simple linear regression with KDE and 95% confidence interval in blue shading along best fit line) between select predictors and predictands. The figure axes are labeled according to the index numbers in table 4.

Figure 9. Joint Plot of Ceiling Height and Surface Wind Speed

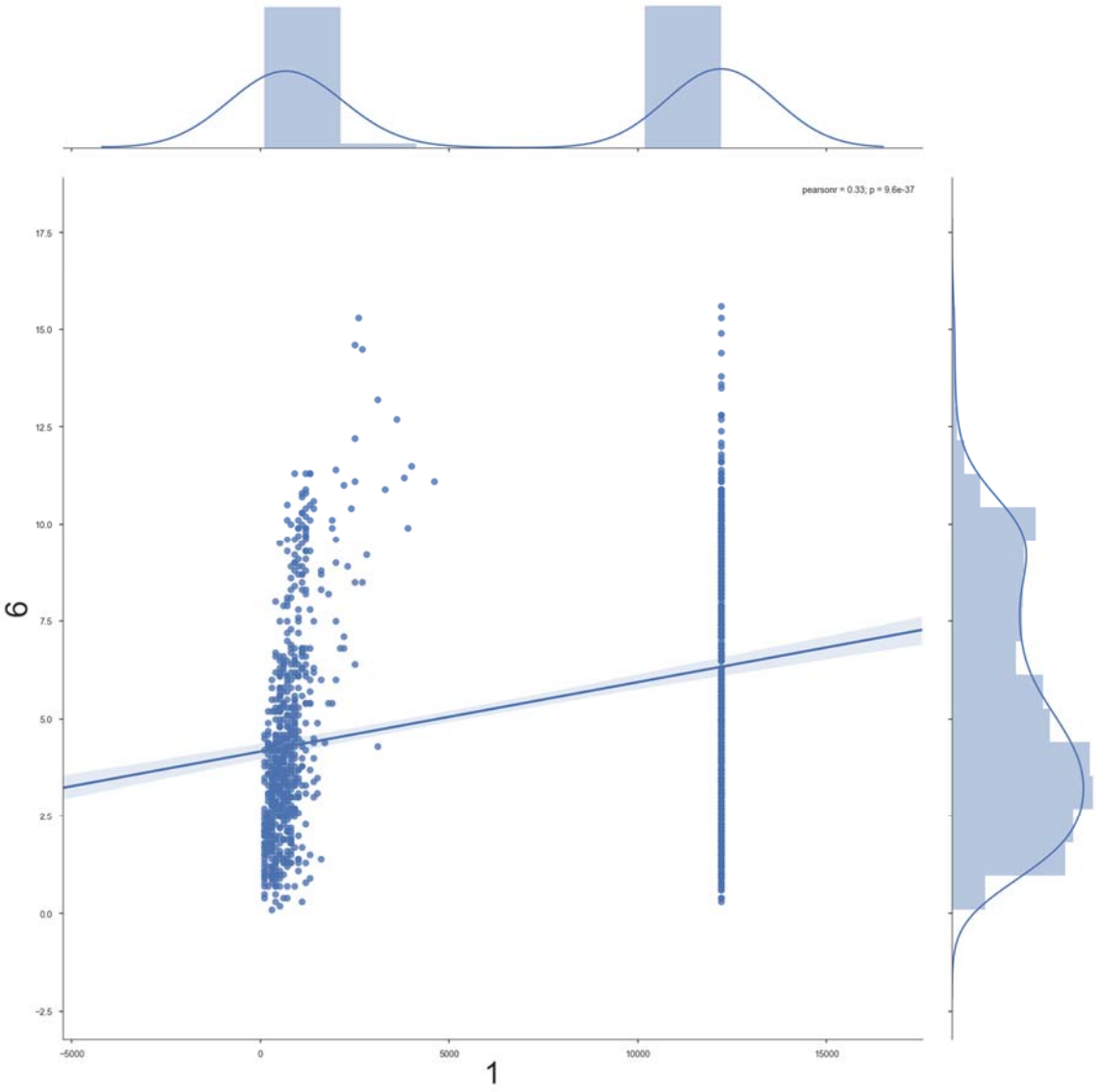


Figure 10. Joint Plot of Visibility and Surface Wind Speed

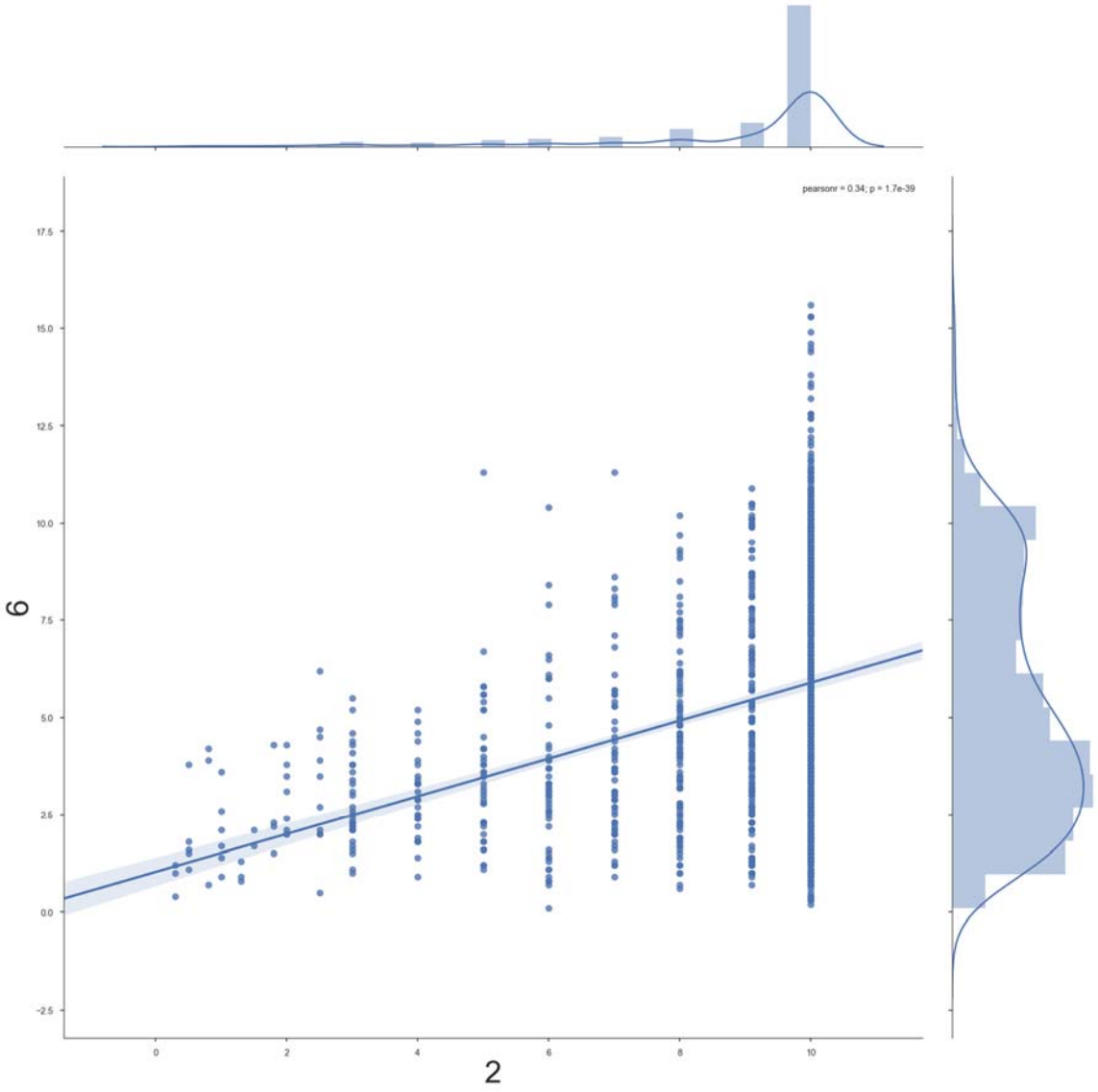


Figure 11. Joint Plot of Visibility and NAM Visibility Forecast

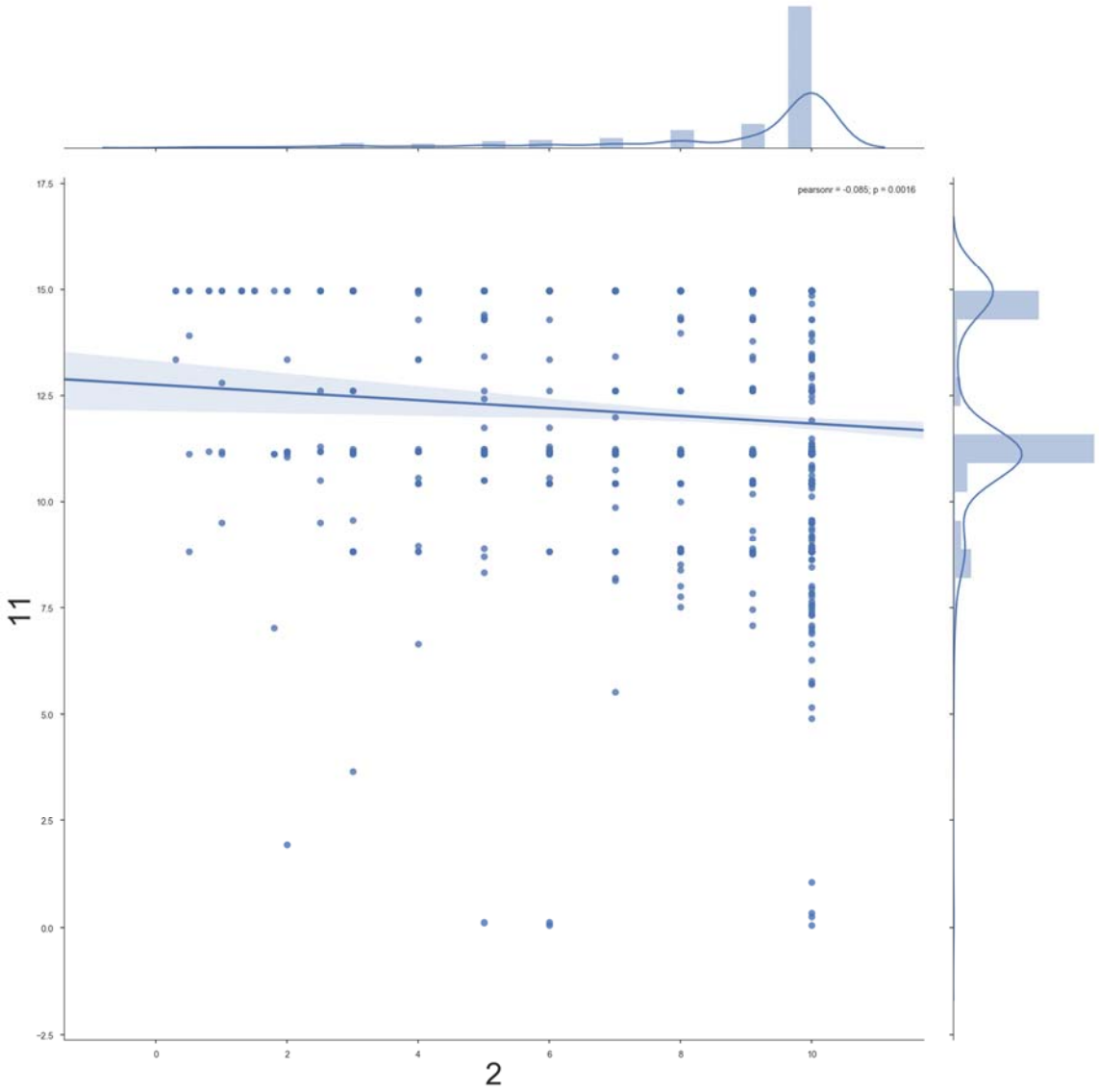
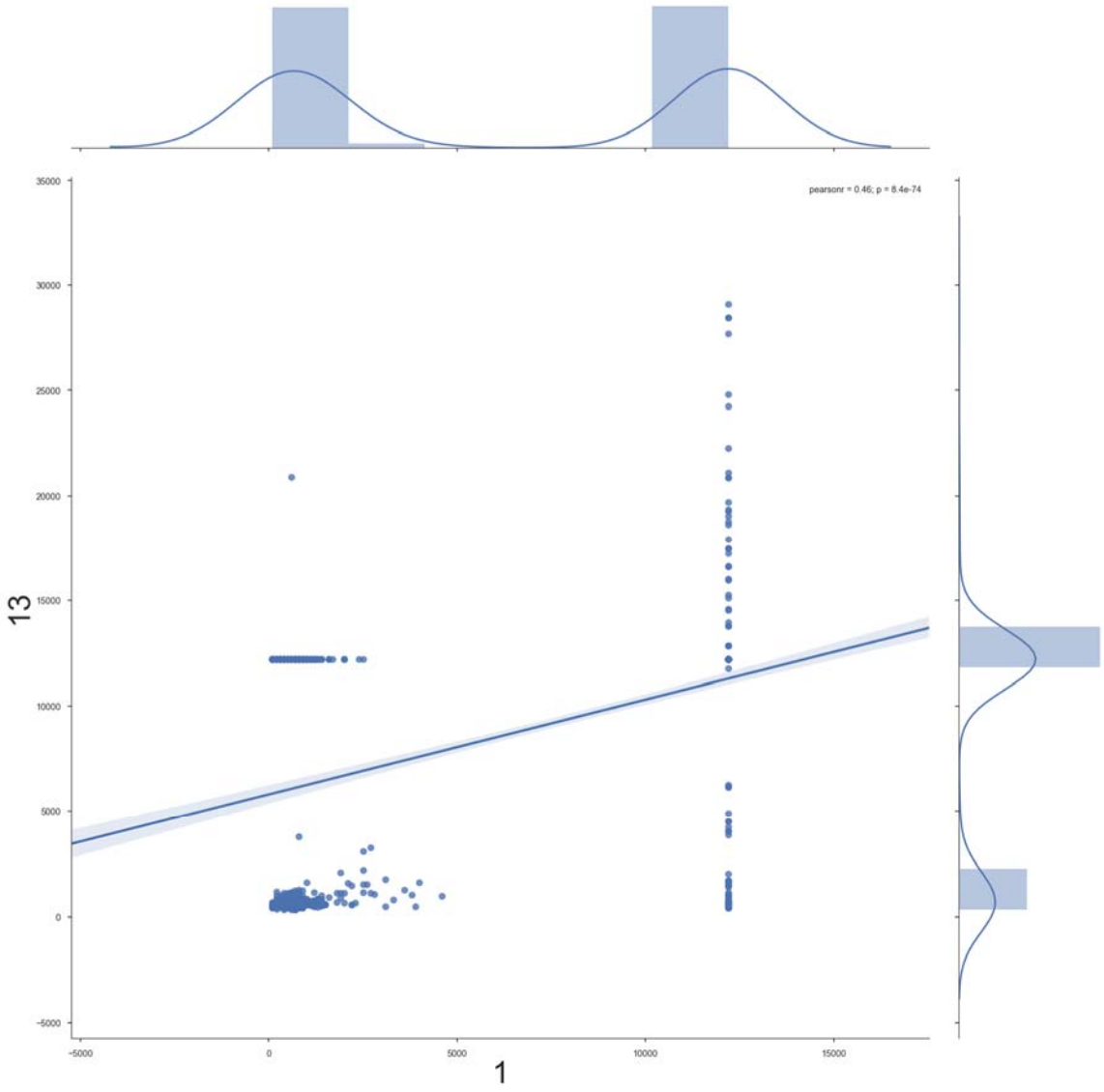


Figure 12. Joint Plot of Ceiling Height and NAM Cloud Base Height Forecast



Comparing each predictand-predictor correlation for the unconditioned data, a number of interesting and important points are apparent. First, figures 9 and 10 show the regressions between ceiling height (r=0.33) and visibility (r=0.34) respectively with surface wind speed. Both appear to be moderately correlated (r=0.3-0.5), but the regression between ceiling height and surface wind speed is problematic because of the bimodal distribution of ceiling height and its artificial cap at 12,200 ft. The regression between visibility and wind speed is similarly impacted by a cap of 10 miles. Next, figures 11 and 12 show the regression between visibility (r=-0.09) and ceiling height (r=0.46) with the NAM visibility and cloud base height forecasts respectively. The NAM visibility forecast is bimodal while the observations are log-normal. This means the correlation between NAM visibility forecast and the observation could be trivial. The predictand and predictor for ceiling height and cloud base are both bimodal which complicates the relationship. The regressions for the remainder of the predictand-predictor relationships were also examined. The bimodal distribution of ceiling height is problematic for all the regressions. Despite these problems, the nowcast system is tested on the unconditioned data to demonstrate its direct application without any intervention via conditioning and clustering.

2. Correlations for Conditioned Data

As previously discussed, the ceiling and visibility observations (predictands) are artificially capped at 12,200 ft and 10 miles respectively. This limits the ability to find relationships between the predictors and predictands. So, conditioning is necessary to focus the investigation on usable data for correlation and regression purposes. Figure 13 displays the histogram and KDE of ceiling height observations (index 1). Figure 14 displays the histogram and KDE of visibility observations (index 2). There is a natural break in the ceiling height data between low (<6000 ft) and high ceilings (>6000 ft). There is not a readily apparent break in the visibility data. So, conditioning was applied to investigate relationships between the predictors and predictands given that the observed ceiling was less than 6000 ft. The correlations before and after conditioning are shown in Table 5.

Figure 13. Histogram and KDE of Ceiling Height Observations

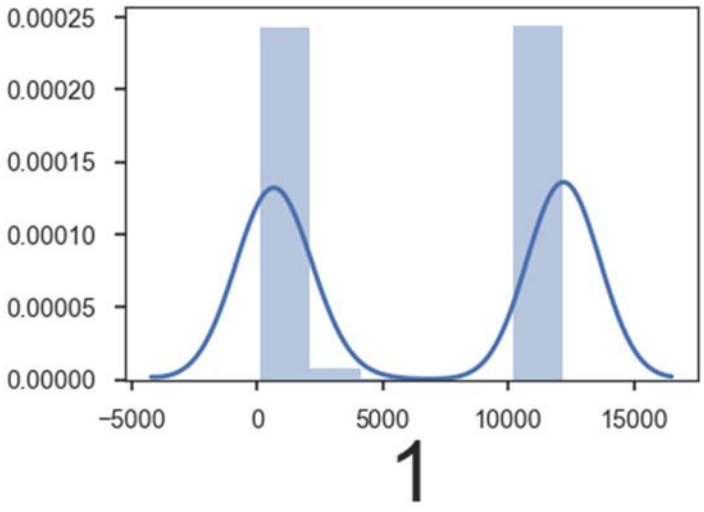


Figure 14. Histogram and KDE of Visibility Observations

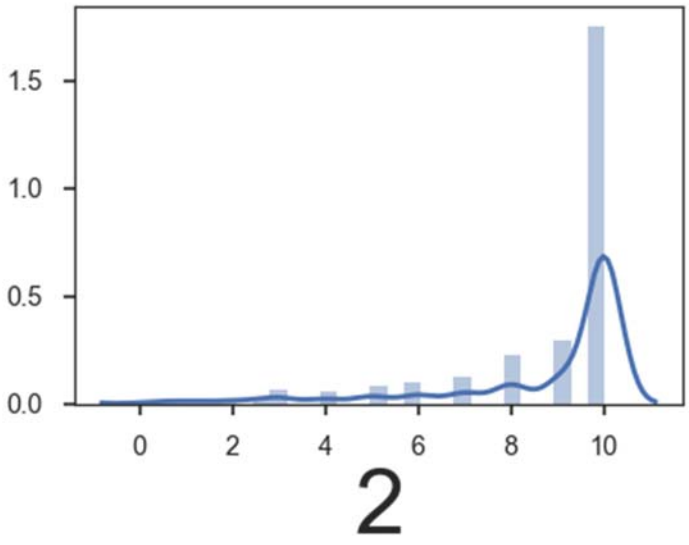


Table 5. Correlation Coefficients for Conditioned and Raw Predictors

Index	Level	Variable Name (units)	Ceiling Conditional r	Raw r	Visibility Conditional r	Raw r
6	sfc	Wind speed (knots)	0.65	0.33	0.36	0.34
11	sfc	Visibility (km)	0.06	0.10	-0.12	-0.09
13	sfc	Cloud base height (m)	-0.11	0.46	-0.08	0.11
39	950	Relative humidity (%)	0.41	-0.35	0.28	0.08
42	975	Relative humidity (%)	0.16	-0.52	0.20	-0.06
43	975	Wind speed (m/s)	0.60	0.21	0.34	0.33
46	1000	Relative humidity (%)	-0.09	-0.55	0.03	-0.15
47	1000	Wind speed (m/s)	0.66	0.31	0.32	0.331

The conditional correlations have changed significantly compared to the raw correlations. Most striking is that wind speeds tend to dominate the correlations after conditioning compared to relative humidity and the raw cloud base height and visibility forecasts. This is reasonable since higher winds could tend to produce higher ceilings and extended visibility ranges if they advect out any marine stratus. The issue with ceiling predictand data being capped is removed, and the model regression will not be biased by the artificial caps. The updated joint plots are shown in figures 15-18. The new regression for ceiling and surface wind speed (figure 15) improved because the artificial height cap at 12,200 ft is removed when we look at low ceilings only. This produced a more useful regression and correlation. Next, the regression for visibility and surface wind speed (figure 16) was not improved because the data was not conditioned on visibility. The new regression for visibility and NAM visibility forecast (figure 17) was similarly not improved. Finally, the regression for ceiling cloud base height (figure 18) was not improved because the predictor is still bimodal. The regressions for the remainder of the predictand-predictor relationships were also examined. The problem of bimodal distribution of ceiling height was resolved for the other predictors.

Figure 15. Joint Plot of Ceiling Height <6000 ft and Surface Wind Speed

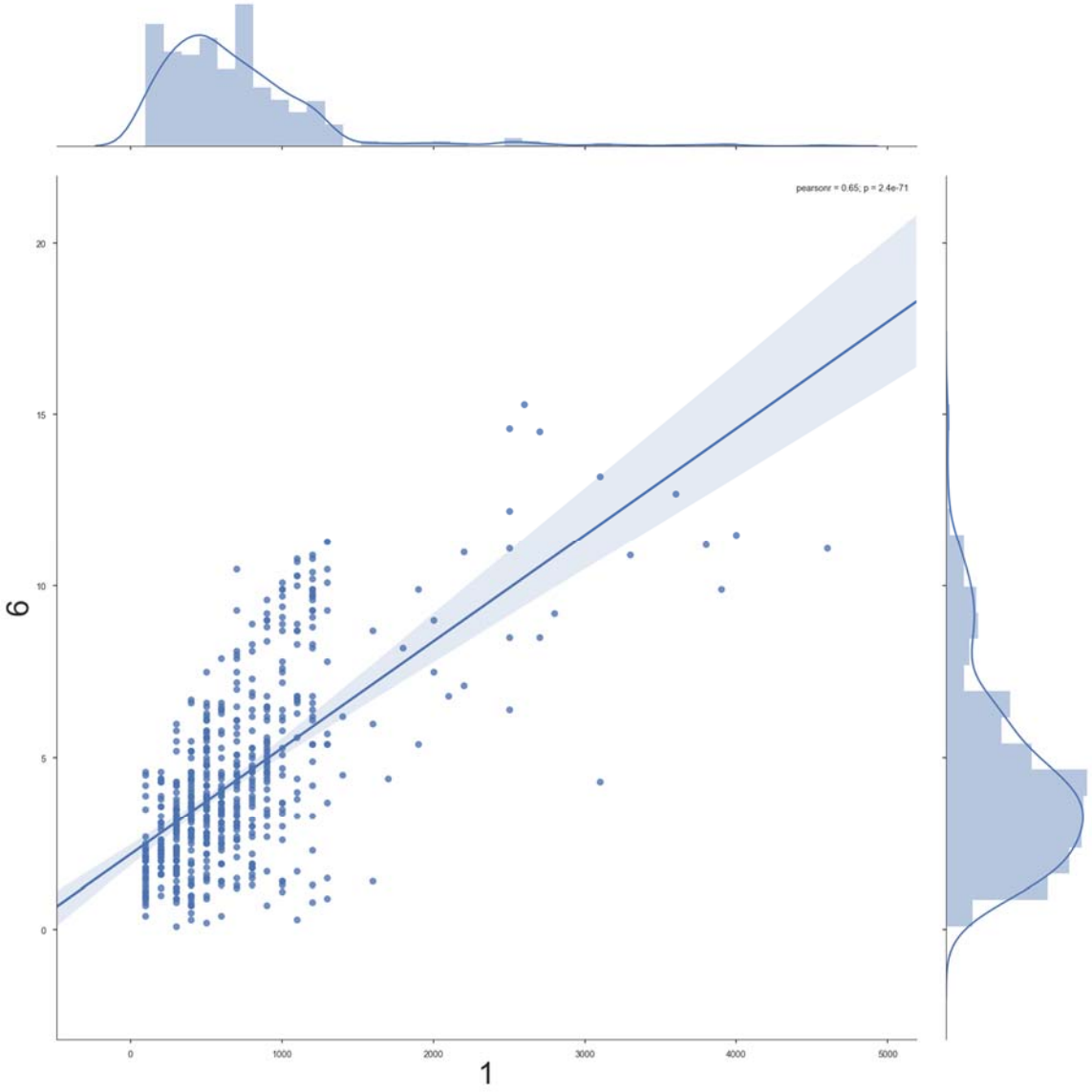


Figure 16. Joint Plot of Visibility (Given Ceiling Height <6000 ft) and Surface Wind Speed

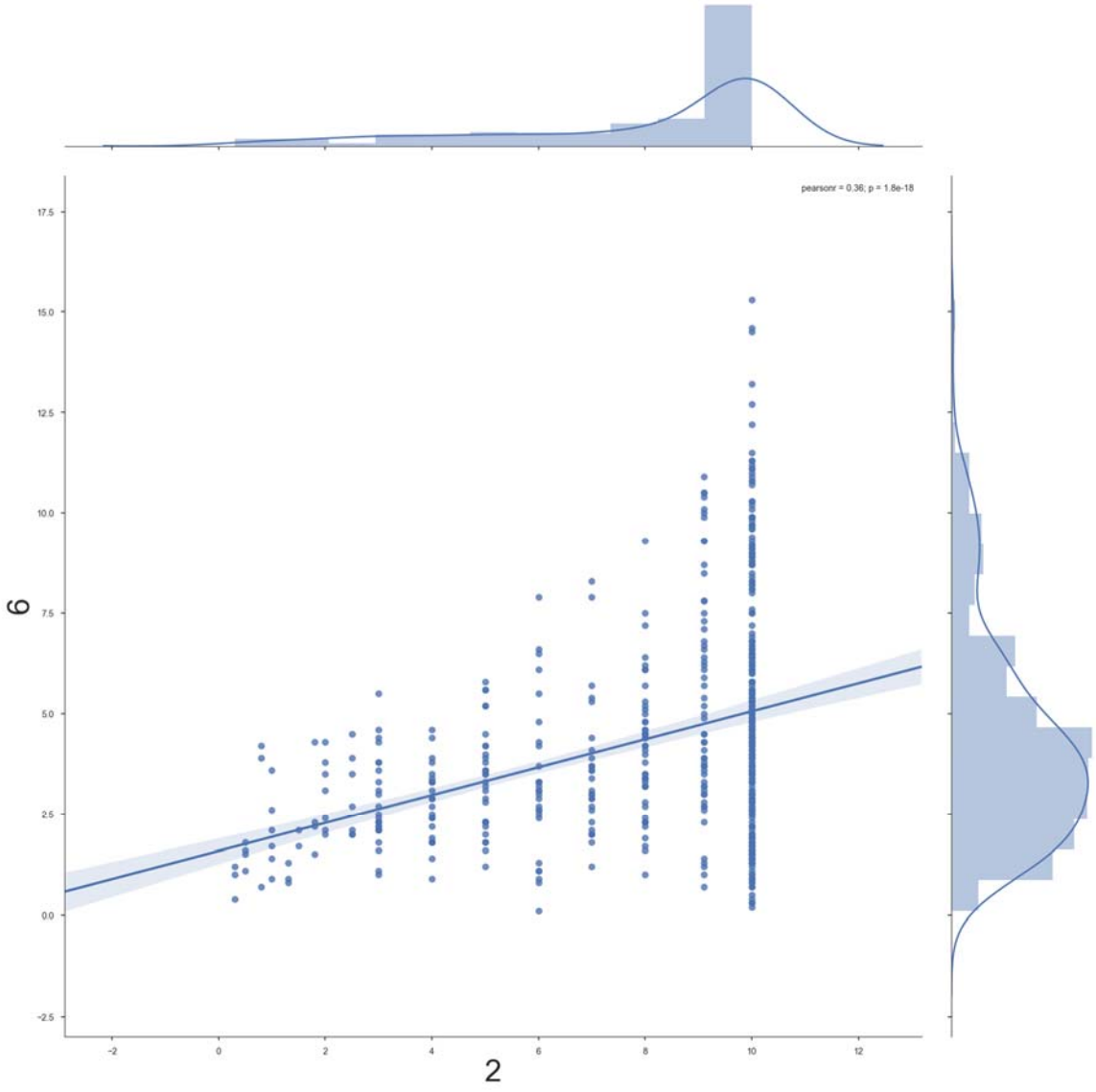


Figure 17. Joint Plot of Visibility (Given Ceiling Height <6000 ft) and NAM Visibility Forecast

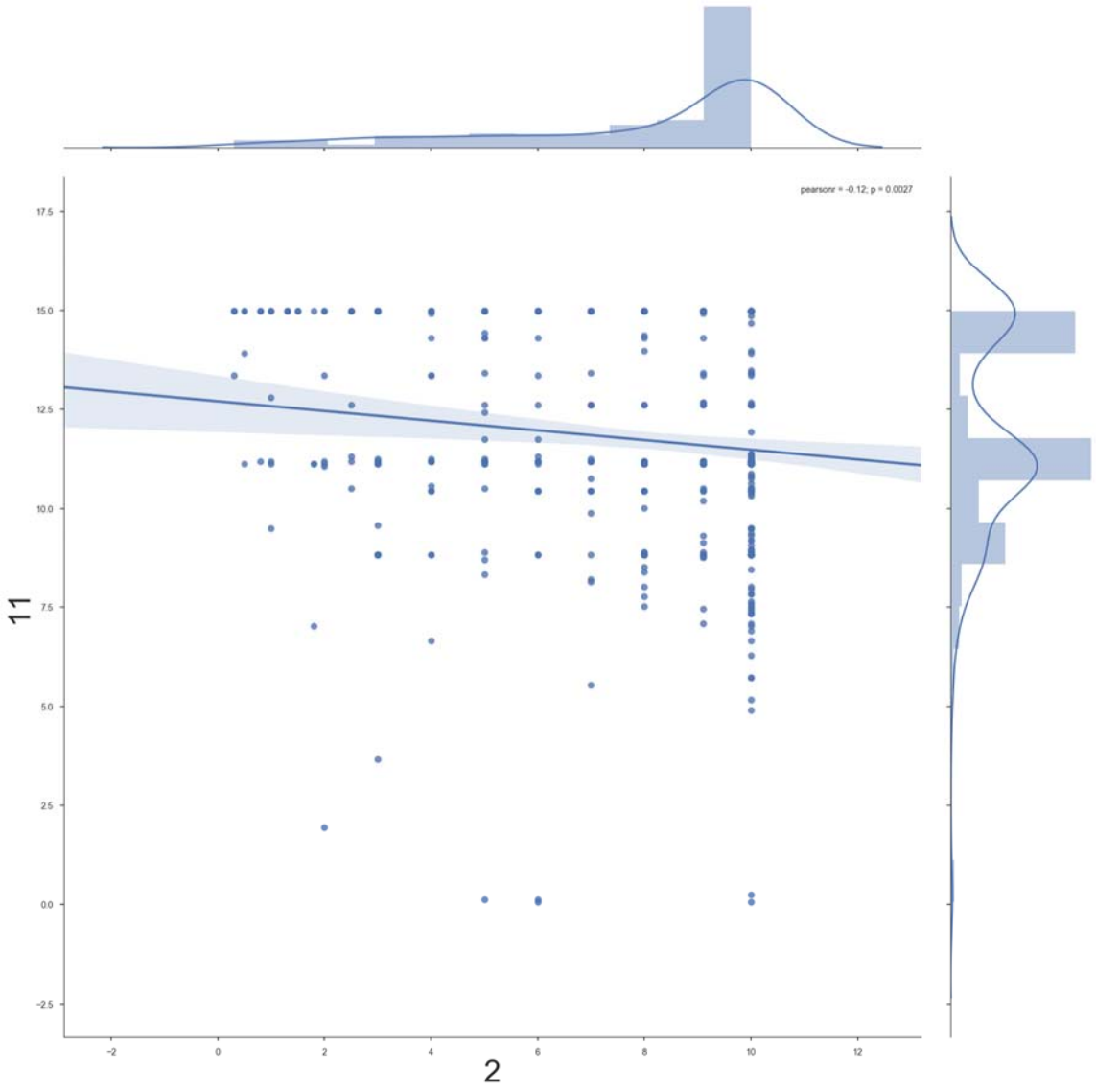
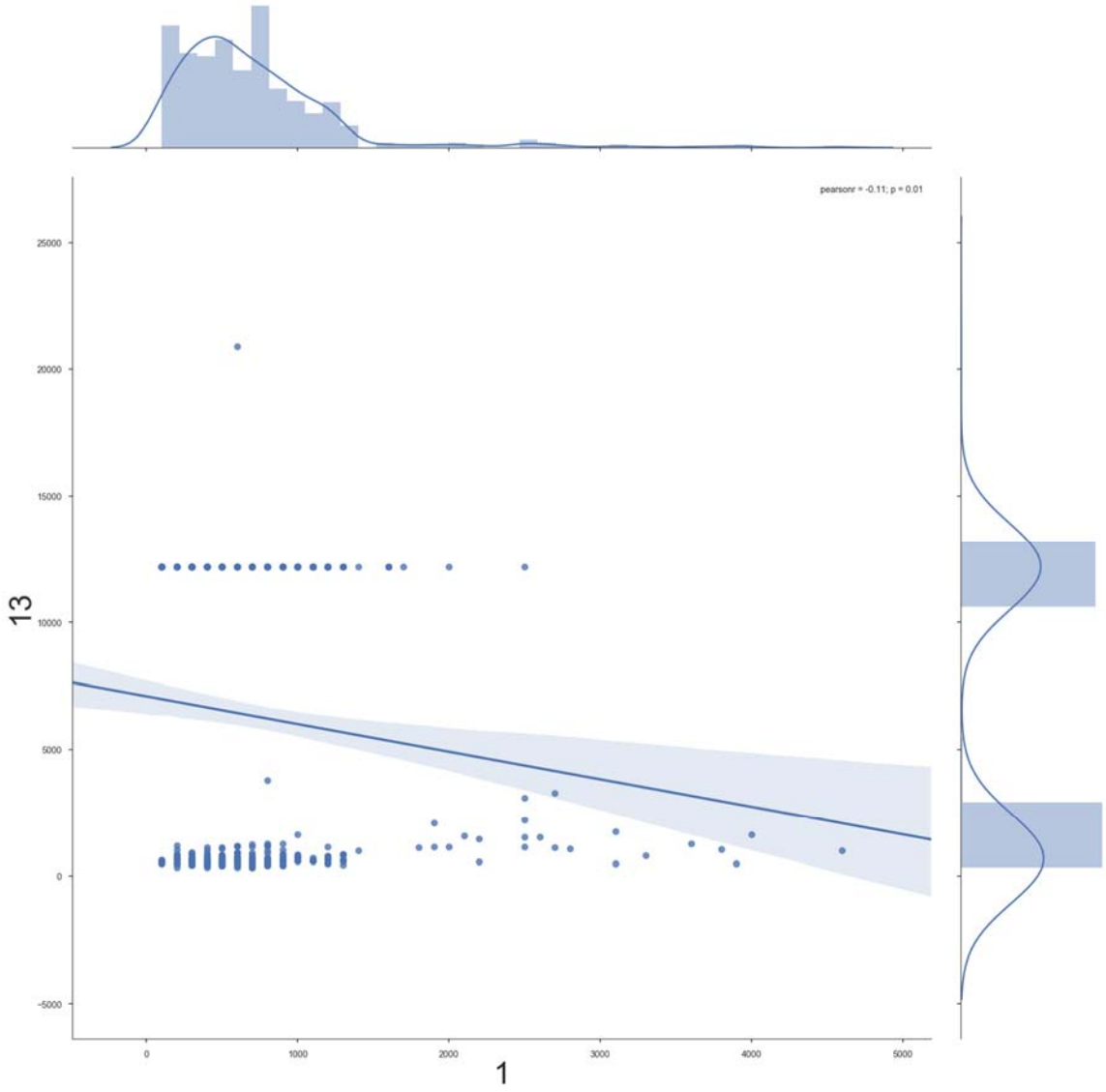


Figure 18. Joint Plot of Ceiling Height <6000 ft and NAM Cloud Base Height Forecast



B. MODEL TRIAL SETTINGS AND METADATA

After investigating the differences in correlations between unconditioned and conditioned data, the following conditions in table 6 were used to test the model.

Table 6. Model Conditions for Each Trial

Run	Predictors	Clustering Method	Conditioning Method
1	Cloud base; visibility	none	none
2	Cloud base; visibility	none	Trained and tested on ceilings <6000ft
3	Cloud base; visibility	K-means	none
4a	Cloud base; visibility	K-means	Trained and tested on ceilings <6000ft
4b	Cloud base; visibility	K-means	Trained on ceilings <6000ft, tested on ceilings >6000ft
4c	Cloud base; visibility	K-means	Trained and tested on ceilings >6000ft
5	surface wind, 950mb relative humidity, 975mb relative humidity, 975mb wind speed, 1000mb relative humidity, and 1000mb wind speed	none	none
6	surface wind, 950mb relative humidity, 975mb relative humidity, 975mb wind speed, 1000mb relative humidity, and 1000mb wind speed	none	Trained and tested on ceilings <6000ft
7	surface wind, 950mb relative humidity, 975mb relative humidity, 975mb wind speed, 1000mb relative humidity, and 1000mb wind speed	K-means	none
8a	surface wind, 950mb relative humidity, 975mb relative humidity, 975mb wind speed, 1000mb relative humidity, and 1000mb wind speed	K-means	Trained and tested on ceilings <6000ft
8b	surface wind, 950mb relative humidity, 975mb relative humidity, 975mb wind speed, 1000mb relative humidity, and 1000mb wind speed	K-means	Trained on ceilings <6000ft, tested on ceilings >6000ft
8c	surface wind, 950mb relative humidity, 975mb relative humidity, 975mb wind speed, 1000mb relative humidity, and 1000mb wind speed	K-means	Trained and tested on ceilings >6000ft

Each combination of predictors was tested: raw, conditioned without clustering, clustered without conditioning, and clustered and conditioned. Moreover, the clustered and conditioned trials were tested on the opposing conditions to investigate false alarm rates. For example, the predictors were trained and tested on ceilings less than 6000 ft, trained on ceilings less than 6000 ft and tested on ceilings greater than 6000 ft, and trained and tested in ceilings greater than 6000 ft. For trials with clustered data, the nowcast model is run on each cluster and errors are calculated separately. Then, the overall errors are calculated for the results from the combined clusters. The predictands for ceiling height and visibility are coupled in the multivariate approach.

Metadata for each trial is listed here in table 7. Train and test sample size refers to the number (hours) of predictand-predictor data sets used to train the model and test the model. The MCMC run time is the total computer time needed to complete the MCMC and BE process. The number of proposals N is the chain length. This number is required to be large to allow for convergence and must be increased if a predictor parameter does not converge ($r<<1$). The burn value refers the point specified by the user to begin sampling the distributions to construct the PPDs. The burn value must be set to a point after convergence and may need to be adjusted. The number of clusters k refers to the number of clusters chosen using k-means cluster analysis for trials that included clustering. Finally, PPD samples refers to the numbers of samples taken from the distribution to construct the PPDs

Table 7. Trial Metadata

Run	Train	Test	Run Time	N	Burn Value	k	PPD samples
1	1183	196	4 min	1e6	5e5	N/A	1e4
2	359	85	3.65 min	1e6	5e5	N/A	1e4
3	874	196	4 min	1e6	5e5	4	1e4
4a	491	125	3.4 min	1e6	5e5	4	1e4
4b	610	71	3.5 min	1e6	5e5	2	1e4
4c	336	71	3.2 min	1e6	5e5	3	1e4
5	1183	196	4 min	1e6	5e5	N/A	1e4
6	359	85	3.55 min	1e6	5e5	N/A	1e4
7	1183	196	4 min	1e6	5e5	4	1e4
8a	573	125	4 min	1e6	5e5	3	1e4
8b	573	71	3.2 min	1e6	5e5	3	1e4
8c	287	71	3 min	1e6	5e5	3	1e4

C. RESULTS

1. Error Comparison

The following tables display the results for the various trials listed in table 6. First, ME, MAE and RMSE are reported for the two-predictor scenario for both ceiling and visibility in tables 8-13. Next, the same is reported for the 6-predictor scenario in tables 14-19. The first column "None" refers to the disuse of both conditioning and clustering. The next column refers to events conditioned on low ceilings only. The third column refers to errors using clustering only. Finally, the last column lists the errors when both conditioning and clustering are used. For trials with clustering, errors are for the overall (combined) test sample.

Table 8. Mean Error (ft) for Ceiling (2 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	2242	2242	1817	5674
NPS Mean	3156	2802	2488	5.25
NPS Median	-2161	2358	1991	-98.5

Table 9. Mean Absolute Error (ft) for Ceiling (2 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	3301	3301	3037	5964
NPS Mean	3864	3562	3353	459
NPS Median	4000	3309	3048	461

Table 10. Root Mean Square Error (ft) for Ceiling (2 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	6007	6007	5372	6402
NPS Mean	6001	5984	5292	959
NPS Median	5337	5993	5314	939

Table 11. Mean Error (miles) for Visibility (2 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	3.19	3.19	3.28	3.72
NPS Mean	0.43	0.75	-0.10	0.86
NPS Median	-0.43	0.44	0.88	0.62

Table 12. Mean Absolute Error (miles) for Visibility (2 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	3.50	3.50	3.53	4.07
NPS Mean	1.48	1.42	1.51	1.76
NPS Median	1.76	1.49	1.53	1.75

Table 13. Root Mean Square Error (miles) for Visibility (2 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	4.46	4.47	4.47	5.01
NPS Mean	2.09	2.20	2.22	2.46
NPS Median	2.10	2.20	2.17	2.39

Table 14. Mean Error (ft) for Ceiling (6 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	2242	4169	2550	4614
NPS Mean	587	3318	1202	-66.8
NPS Median	-2234	1670	-344	-145

Table 15. Mean Absolute Error (ft) for Ceiling (6 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	3301	4523	3587	4868
NPS Mean	3765	3322	3986	263
NPS Median	3640	1961	3716	293

Table 16. Root Mean Square Error (ft) for Ceiling (6 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	6007	6999	6245	6522
NPS Mean	5187	5216	5762	338
NPS Median	5371	3883	5396	364

Table 17. Mean Error (miles) for Visibility (6 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	3.19	3.66	3.22	3.44
NPS Mean	0.07	0.24	0.64	0.29
NPS Median	-0.48	-0.45	0.08	-0.58

Table 18. Mean Absolute Error (miles) for Visibility (6 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	3.50	4.09	3.50	3.79
NPS Mean	1.66	1.85	1.58	1.94
NPS Median	1.86	2.07	1.83	2.28

Table 19. Root Mean Square Error (miles) for Visibility (6 Predictors)

	None	Conditioned	Clustered	Clustered and Conditioned
NAM	4.46	5.09	4.23	4.63
NPS Mean	2.35	2.48	2.31	2.52
NPS Median	2.39	2.48	2.49	2.77

2. Interpretation

First, considering the two-predictor scheme using NAM cloud base height and visibility forecasts to forecast ceiling height, tables 8-10 show that ME, MAE and RMSE are significantly reduced only when both clustering and conditioning are applied (run 4a). This result makes sense because the bimodal ceiling predictand behavior is removed by conditioning. Moreover, clustering reduces variance by finding the appropriate groupings of data. Clustering alone did not improve forecasts most likely because the bimodal predictand distribution resulted in many incorrect cluster assignments thereby increasing variance.

For visibility, the post-processing of raw data (run 1:no conditioning or clustering) produced the lowest values of ME, MAE, and RMSE for the two-predictor scheme (see tables 11-13). This is most likely because the predictand distribution is log-normal and therefore does not pose a significant problem compared to the ceiling height distribution. All methods improve the visibility forecast compared to the raw NAM visibility forecast, but conditioning and/or clustering were not advantageous over the use

of the raw data. In most cases, the NPS mean forecast performed better (smaller errors) than the median forecast.

Next, considering the 6-predictor scheme using NAM cloud base height and visibility forecasts to forecast ceiling height, tables 14-16 show that ME, MAE and RMSE are also significantly reduced only when both clustering and conditioning are applied (run 8a). This result makes sense for the same reasons as before. For visibility, the post-processing of clustered data (run 7) produced the lowest values of ME and MAE for the 6-predictor scheme (see tables17-19). The smallest RMSE was produced using the raw data. This is most likely because the predictand distribution is log-normal and therefore does not pose a significant problem compared to the ceiling height distribution. All methods improve the visibility forecast compared to the raw NAM visibility forecast, but clustering was usually advantageous over the use of the raw data. In most cases, the NPS mean forecast performed better than the median forecast.

Comparing the two-predictor and the 6-predictor scheme for ceiling height forecasts, all error calculations were improved most significantly by the 6-predictor scheme. For example, the MAE for the raw NAM ceiling height forecast was over 4,000 ft. The best performing NPS forecast reduced MAE to 459 ft for the 2-predictor scheme and 263 ft for the 6-predictor scheme. The RMSE for the raw NAM ceiling height forecast was over 6,000 ft. The best performing NPS forecast reduced RMSE to 939 ft for the 2-predictor scheme and 338 ft for the 6-predictor scheme.

Next, comparing the two-predictor and the 6-predictor scheme for visibility forecasts, all error calculations were improved most significantly by the 2-predictor scheme. For example, the MAE for the raw NAM visibility forecast was over 3 miles. The best performing NPS forecast reduced MAE to 1.42 miles for the 2-predictor scheme and 1.58 miles for the 6-predictor scheme. The RMSE for the raw NAM visibility forecast was over 4 miles. The best performing NPS forecast reduced RMSE to 2.09 miles for the 2-predictor scheme and 2.31 miles for the 6-predictor scheme.

Thus, in general, the 6-predictor NPS mean forecast using conditioning and clustering resulted in the smallest errors for ceiling height forecasts. The two-predictor

NPS mean forecast without conditioning or clustering resulted in the smallest errors for visibility forecasts. When examining only low ceiling test days (conditioned data), the NPS nowcast using physical predictors dramatically improved upon the raw NAM forecasts. However, these results must be taken in context. This model trial was trained on cases when the ceiling height observation (predictand) was less than 6000 ft, and tested on cases when the verifying observation was also less than 6000 ft. The model was not given this information, but was it was tuned for low ceiling events. So, while the nowcast system greatly reduced error for low ceiling forecasts, the model must be tested on the opposite condition to test for false alarms. Run 8b tested the model tuned for low ceilings on cases when the verifying observation was greater than 6000 ft. This resulted in the NPS model dramatically increasing ceiling height error by over 100% for this scenario. This means that the model tuned for low ceilings cannot be applied to cases of high ceilings. Operationally this is significant, because if this model were to be operational as-is, the forecaster would need to make an inference about whether or not the real outcome of the ceiling forecast will be less than or greater than 6000 ft. This inference can be made practically by assuming persistence (a recent observation indicates a ceiling less than 6000 ft) or using remote sensing such as satellite imagery to confirm the presence of low clouds a priori.

D. CLUSTERING RESULTS AND MCMC DIAGNOSTICS

For each trial, diagnostics were collected to examine the k-means clustering results and MCMC and BE performance. A sample of the diagnostics for trials 8a, the best performing scheme, are included here for review. First, the silhouette score distribution and analysis for trial 8a (6 predictors, conditioned and clustered) are shown in figures 19 and 20. Based on the clustering analysis, n=3 clusters was chosen because it had a relatively high score/low variance and minimal cluster misassignment (negative scores). Other clustering combinations either lower or similar scores but with more misassignment. Refer to the methods section for cluster analysis discussion.

Figure 19. Silhouette Scores for Trial 8a

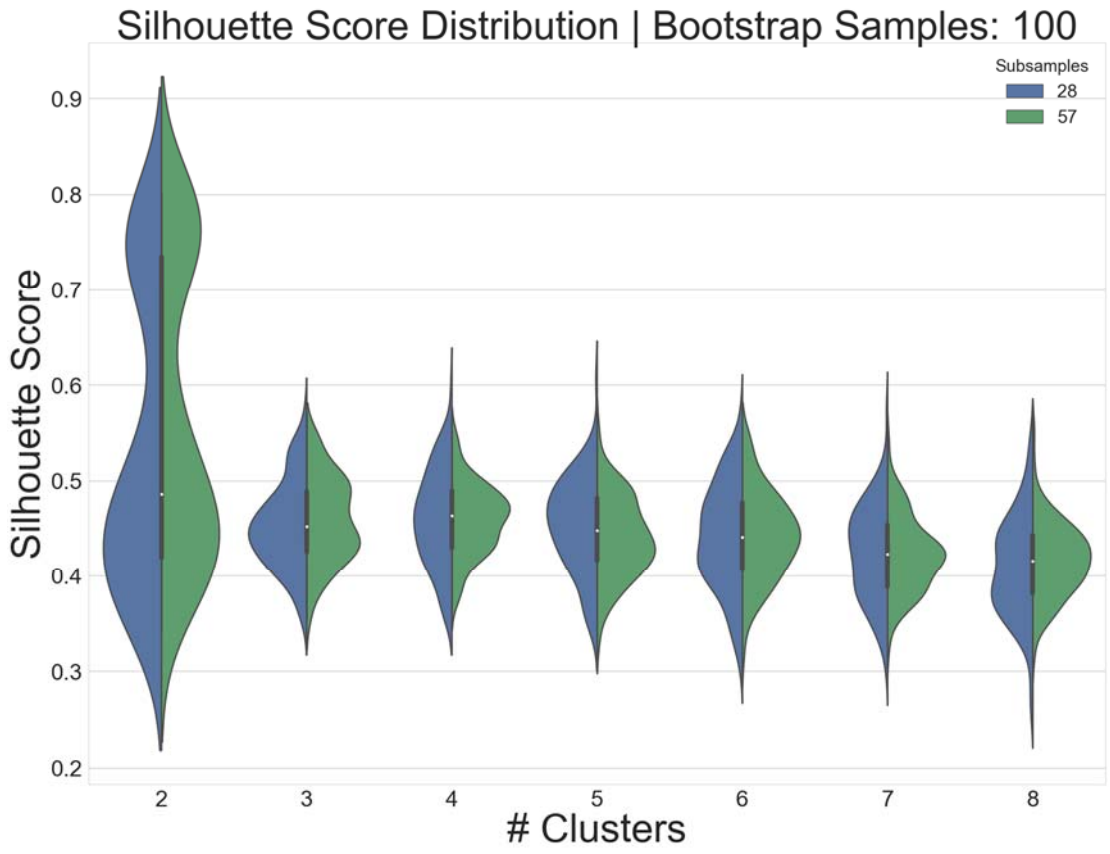
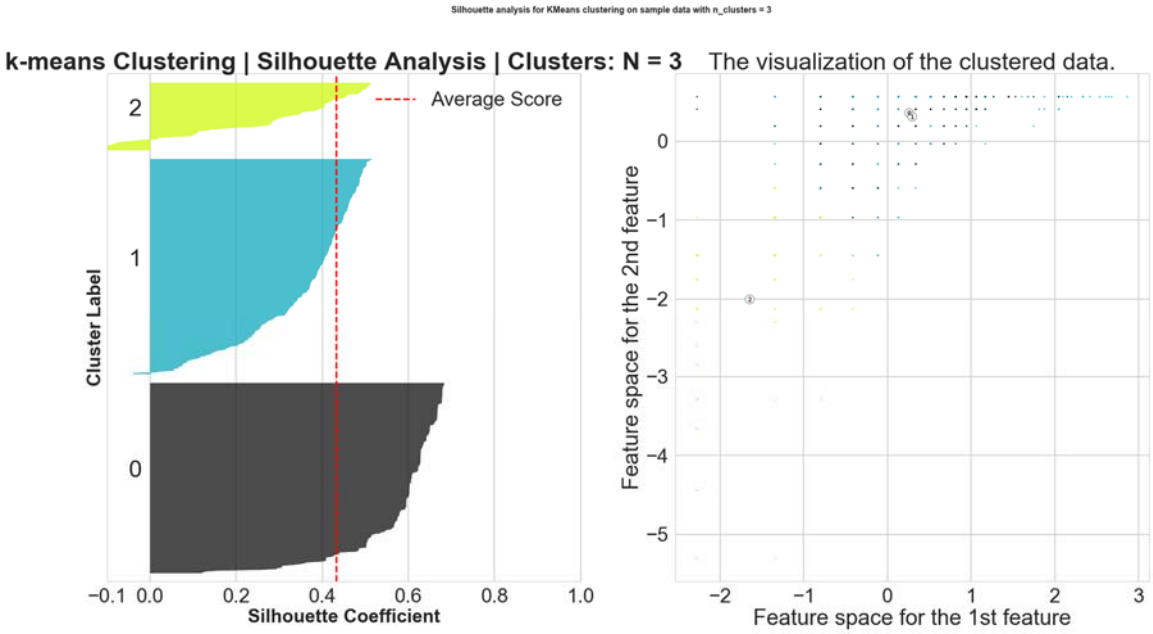


Figure 20. Silhouette Analysis for Trial 8a



Next, figure 21 displays the MCMC chain trace for θ_4 (β_4) from trial 8a cluster 2, which is the regression coefficient distribution for the NAM cloud base height predictor and ceiling height predictand. Cluster 2 was chosen for review because it had the greatest number of training and test data points. From before, the correlation coefficient between the cloud base height predictor and ceiling height predictand was only r=.11. After post-processing, the Bayesian estimation of the regression weight is around .5 (the mean of the distribution as identified by the thin blue line). Relative to the other predictor (NAM visibility forecast), the cloud base height predictor was a more explanatory predictor. This makes sense because it is reasonable to assume that as the NAM cloud base height forecast increases, so should the ceiling height.

Figure 21. MCMC Chain Trace of β4

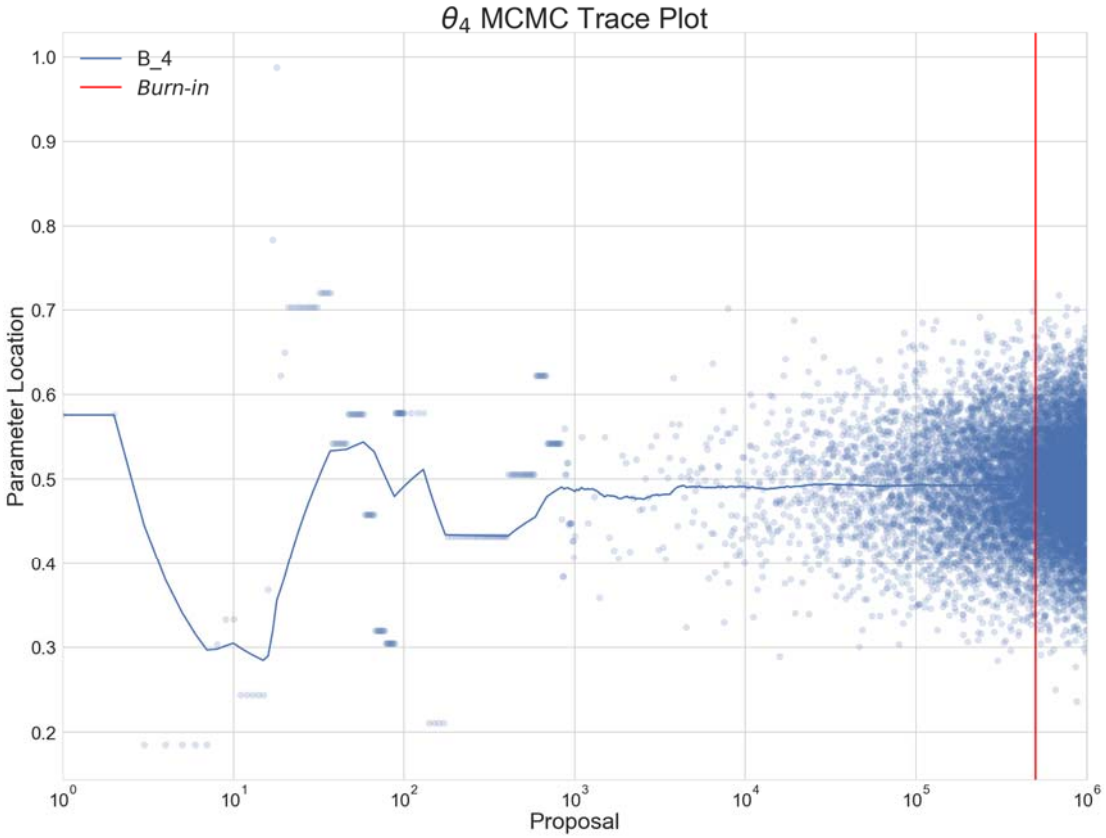
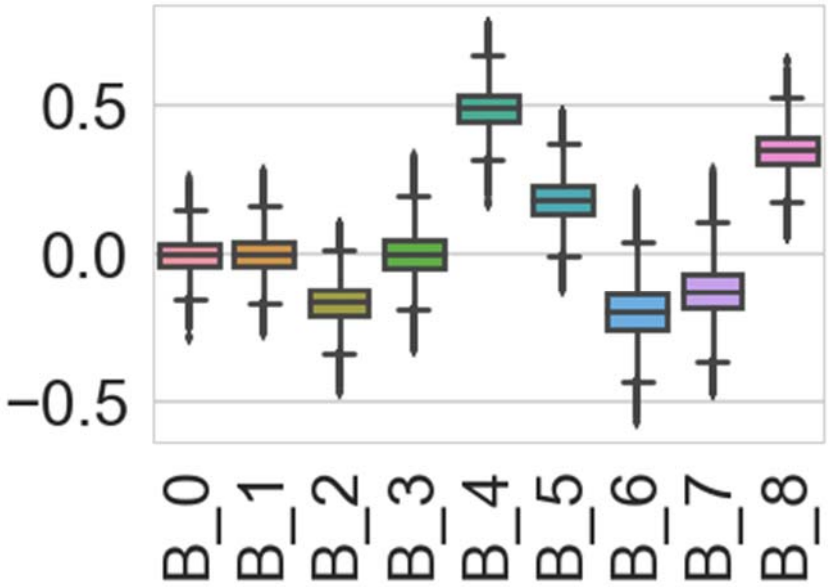


Figure 22 shows the model parameter boxplot for trial 8a cluster 2, which summarizes the distributions for all parameters. Each box plot corresponds to the intercepts and regression parameters for the ceiling height and visibility GLM respectively. So, β_0 and β_1 are the intercepts for ceilings height and visibility, β_2 and β_3 are the coefficients for the NAM visibility forecast predictor, β_4 and β_5 are the coefficients for the NAM cloud base height forecast predictor, and β_6-β_8 are the parameter estimates of the covariance matrix. A similar figure was produced for all trials and all clusters to evaluate the relative differences in predictor regression coefficient distributions. The results indicate that predictors with high correlations coefficients were the best explanatory variables for predictions

Figure 22. Boxplot of Coefficient Distributions for Trial 4a Cluster



E. SUMMARY OF RESULTS AND INTERPRETATION

1. Run 8a Ceiling Height Forecasts

Tables 20-22 display a sample set of ceiling forecasts for Run 8a per flight category: LIFR, IFR and MVFR. Forecasts for VFR are not shown because they correspond to Run 8b. The first column is the date time group (DTG). They are not in DTG order because they are sorted by flight category. The second column is the ceiling height observation. The third column is the corresponding flight category for scoring purposes. The next 2 columns are the NAM cloud base height forecast and the corresponding flight category. The final 2 columns are the NPS mean ceiling height forecast and the corresponding flight category. Only the NPS mean is shown because it is the best performing forecast compared to the median in this case.

Table 20. LIFR Ceiling Forecasts

DTG	Cig Ob	Flight Category	NAM	Flight Category	NPS mean	Flight Category
17081912	400	LIFR	649.82	IFR	568.68	IFR
17082805	400	LIFR	11941.16	VFR	282.78	LIFR
17082806	300	LIFR	11941.16	VFR	317.72	LIFR
17082807	400	LIFR	11941.16	VFR	302.39	LIFR
17082808	300	LIFR	11941.16	VFR	273.84	LIFR

Table 21. IFR Ceiling Forecasts

DTG	Cig Ob	Flight Category	NAM	Flight Category	NPS mean	Flight Category
17081412	500	IFR	617.01	IFR	845.93	IFR
17081414	500	IFR	249.56	LIFR	693.64	IFR
17081415	600	IFR	256.12	LIFR	670.63	IFR
17081416	700	IFR	341.42	LIFR	695.64	IFR
17081417	800	IFR	396.54	LIFR	736.95	IFR

Table 22. MVFR Ceiling Forecasts

DTG	Cig Ob	Flight Category	NAM	Flight Category	NPS mean	Flight Category
17081421	1600	MVFR	661.6319	IFR	1697.8408	MVFR
17081422	1900	MVFR	640.63452	IFR	1922.2985	MVFR
17081423	1400	MVFR	523.83662	IFR	2016.5738	MVFR
17081600	2000	MVFR	11941.163	VFR	1911.4016	MVFR
17081601	2400	MVFR	11941.163	VFR	1914.3565	MVFR

2. Run 8a Visibility Forecasts

Tables 23-26 display a sample set of visibility forecasts for Run 8a per flight category: LIFR, IFR MVFR, and VFR. The first column is the date time group (DTG). They are not in DTG order because they are sorted by flight category. The second column is the visibility observation. The third column is the corresponding MOS flight category for scoring purposes. The next 2 columns are the NAM visibility forecast and the

corresponding flight category. The final 2 columns are the NPS mean visibility forecast and the corresponding MOS and flight category. Only the NPS mean is shown because it is the best performing forecast compared to the median in this case.

Table 23. LIFR Visibility Forecasts

DTG	Vis Ob	Flight Category	NAM	Flight Category	NPS mean	Flight Category
17082914	0.5	LIFR	8.82	VFR	8.12	VFR

Table 24. IFR Visibility Forecasts

DTG	Vis Ob	Flight Category	NAM	Flight Category	NPS mean	Flight Category
17081413	2.5	IFR	14.98	VFR	9.08	VFR
17082314	3	IFR	12.61	VFR	8.86	VFR
17082315	2.5	IFR	9.51	VFR	8.62	VFR
17082316	3	IFR	12.61	VFR	8.30	VFR
17082912	3	IFR	14.98	VFR	8.40	VFR

Table 25. MVFR Visibility Forecasts

DTG	Vis Ob	Flight Category	NAM	Flight Category	NPS mean	Flight Category
17081414	5	MVFR	10.50	VFR	9.15	VFR
17081415	5	MVFR	11.25	VFR	8.97	VFR
17081417	5	MVFR	14.98	VFR	8.96	VFR
17082015	5	MVFR	14.98	VFR	8.15	VFR
17082318	5	MVFR	14.98	VFR	7.82	VFR

Table 26. VFR Visibility Forecasts

DTG	Vis Ob	Flight Category	NAM	Flight Category	NPS mean	Flight Category
17081412	7	VFR	14.98	VFR	9.87	VFR
17081416	6	VFR	10.56	VFR	8.86	VFR
17081815	10	VFR	14.98	VFR	8.36	VFR
17081816	7	VFR	14.98	VFR	7.67	VFR
17081905	10	VFR	11.12	VFR	8.32	VFR

a. Run 8a scoring

When reviewing the scoring, it is important to remember that Run 8a forecasts are conditional upon low ceilings. Tables 27 and 28 display the overall percent correct (PC) and skill scores (SS) for the full set ceiling and visibility forecasts for Run 8a. In terms of skill scores, the NPS mean ceiling height forecast greatly reduced RMSE by almost 100% compared to the raw NAM cloud base height forecast. The NPS mean forecast reduced error by about 50% for visibility. However, this improvement did not translate to forecast skill improvement. This is because a visibility forecast over 10 miles is penalized for distance over 10 miles. So, if the observation is 10 miles, then any forecast for both NAM and NPS is penalized if it is greater than 10 miles even though it is not operationally wrong. The raw NAM visibility forecast was correct most of the time, even though it tended to miss the low visibility cases. The NPS model was not able to do any better than the raw forecast because the predictand data is log-normal and highly skewed left. The NPS model assumes a normal likelihood function and was not able to catch the extreme events any better than the pre-processed forecast.

Table 27. Run 8a Skill Scores

	Ceiling	Visibility
NPS Mean	0.94	0.46
NPS Median	0.94	0.40

Table 28. Run 8a Percent Correct by Flight Category

	Ceiling	Visibility
NAM	0.19	0.82
NPS Mean	0.62	0.82
NPS Median	0.45	0.74

Focusing on ceiling height forecasts, table 29 displays the percent correct for the full set of NAM and NPS ceiling height forecasts by flight category. Forecast skill was greatly improved in the LIFR and MVFR flight categories and modestly improved in the IFR category. The RMSE skill scores for the NPS forecast are also displayed in table 30. Across the board, the NPS forecasts reduced error by 90 to 96% for low ceiling events (LIFR, IFR, and MVFR).

Table 29. Run 8a Ceiling Percent Correct by Flight Category

Ceiling	LIFR	IFR	MVFR
NAM	0.43	0.19	0.00
NPS mean	0.50	0.79	0.4
NPS median	0.48	0.48	0.33

Table 30. Run 8a Ceiling Skill Scores by Flight Category

Ceiling	LIFR	IFR	MVFR
NPS mean	0.96	0.96	0.91
NPS median	0.96	0.96	0.90

Next, contingencies tables for NAM and NPS forecasts per flight category are listed in table 31 and 32 respectively. The marginals are the check-sums to ensure all forecasts are counted.

Table 31. Run 8a Contingency Tables by Flight Category for NAM Cloud Base Height Forecast

NAM		Observed		
		LIFR	Other	Marginals
Forecasted	LIFR	12	48	60
	other	16	49	65
	Marginals	28	97	125

		Observed		
		IFR	Other	Marginals
Forecasted	IFR	12	6	18
	other	50	57	107
	Marginals	62	63	125

		Observed		
		MVFR	Other	Marginals
Forecasted	MVFR	0	0	0
	other	125	0	125
	Marginals	125	0	125

Table 32. Run 8a Contingency Tables for NPS Mean Ceiling Height Forecasts

NPS Mean		Observed		
		LIFR	Other	Marginals
Forecasted	LIFR	14	6	20
	Other	14	91	105
	Marginals	28	97	125

		Observed		
		IFR	Other	Marginals
Forecasted	IFR	48	35	83
	Other	13	29	42
	Marginals	61	64	125

		Observed		
		MVFR	Other	Marginals
Forecast	MVFR	14	8	22
	Other	21	82	103
	Marginals	35	90	125

From the contingency tables, forecast metrics are calculated in table 33.

Table 33. Run 8a Contingency Table Metrics

Flight Category	Metric	NAM	NPS Mean	Difference
LIFR	PC	0.49	0.84	0.35
	HR	0.43	0.5	0.07
	FAR	0.80	0.3	-0.50
	TS	0.16	0.41	0.25
	Bias	2.14	1.36	-0.78
	HSS	-0.05	0.49	0.54
IFR		NAM	NPS Mean	
	PC	0.55	0.62	0.07
	HR	0.19	0.79	0.60
	FAR	0.33	0.42	0.09
	TS	0.18	0.5	0.32
	Bias	0.29	1.36	1.07
	HSS	0.10	0.24	0.14
MVFR		NAM	NPS Mean	
	PC	0.00	0.77	0.77
	HR	0.00	0.4	0.40
	FAR	0.00	0.36	0.36
	TS	0.00	0.33	0.33
	Bias	0.00	0.63	0.63
	HSS	0.00	0.35	0.35

For LIFR conditions, the NPS forecast improved PC by 35%, HR by 7%, decreased FAR by 50%, increased TS by 25%, decreased bias (over forecasted less often) and increased HSS by 54%. For IFR conditions, the NPS forecast improved PC by 7%, HR by 60%, slightly increased FAR by 9%, increased TS by 32%, and increased HSS by 14%. Bias did increase significantly for the NPS mean, but the magnitude it over forecasts (1.36) is less than the magnitude NAM under forecasts (.29). For MVFR conditions, NAM did not make any forecasts in this run. Forecasts for LIFR and IFR conditions were improved most dramatically, which is promising because very low ceilings (LIFR indicates ceilings below 500ft AGL and IFR between 500 ft and 1000 ft

AGL) are difficult to forecast and dangerous to operations. Improving forecast skill for these flight categories is most critical to safety and mission success.

b. Probabilistic Forecasts

The aforementioned error statistics and skill scores are important and demonstrate improvement of the raw NAM forecasts. Surely, any system that reduces error and improves skill is potentially useful, especially when it is computationally efficient. However, the most important aspect that the NPS nowcast system offers is the explicit probabilistic information offered through Bayesian inference (Wendt 2017). The key attribute that the nowcast system offers is the ability to generate probabilistic forecasts from deterministic data without using ensemble forecasting (Wendt 2017). For example, table 34 displays one forecast trial's observation, NAM forecasts and NPS mean forecasts for ceiling height and visibility. The NPS mean improves the ceiling height forecast by post-processing the NAM forecast of 375 ft to 569 ft, which is much closer to the truth of 700 ft.

Table 34. Ceiling and Visibility Forecasts for 2017081907

DTG	Ob			NAM			NPS		
	CLG (ft)	Category	VIS (miles)	CLG (ft)	Category	VIS (miles)	CIG (ft)	Category	VIS (miles)
2017082109	700	IFR	8	375	LIFR	10	569	IFR	9

Figures 23 and 24 below display the posterior predictive distributions for the forecast (DTG 2017082109) from Run 8a. The grey shading is the 10,000 discrete forecast samples forming the forecast histogram. The solid gray line is the KDE estimating the continuous probability density function (PDF) from the histogram. The horizontal blue dashed line is the interquartile range (IQR). The vertical black line is the median. The horizontal red line is the 95% high density interval (analogous to confidence interval). The solid blue curve is the cumulative density function (CDF).

Figure 23. Ceiling Height PPD for 2017082109

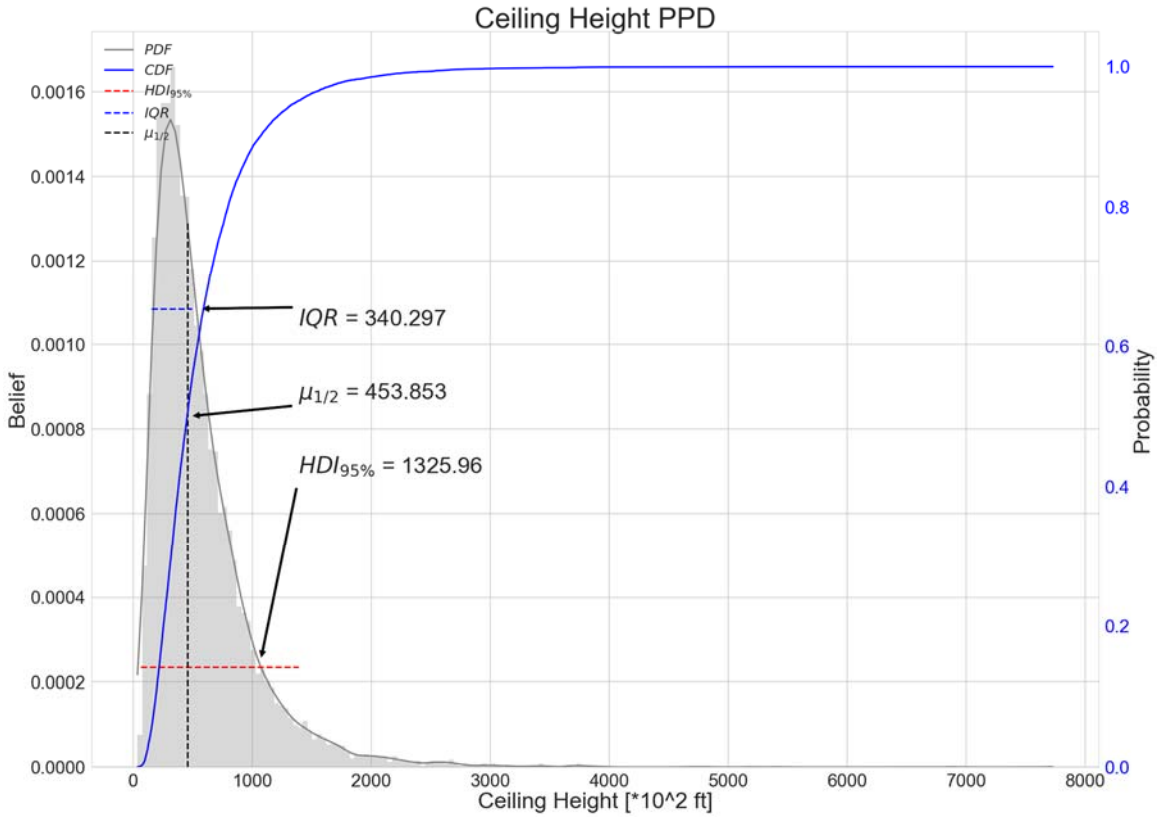


Figure 24. Visibility PPD for 2017082109

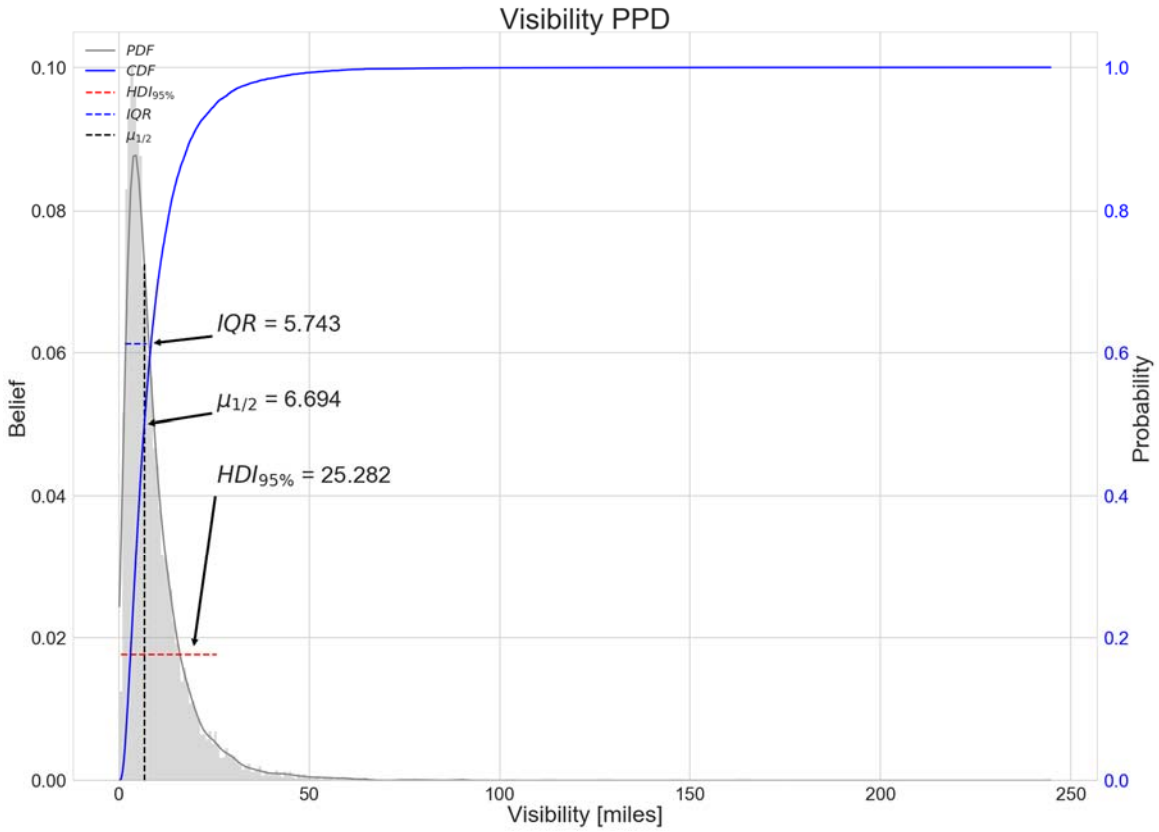


Table 35. High Density Intervals for 2017081907

	68%	80%	95%
Ceiling height (ft)	(124,653)	(96,809)	(65,1391)
Visibility (miles)	(1.1,10.3)	(.87,13.6)	(.43,25.7)

First, from figure 23, most of the density of the ceiling height samples for 2017082109 is visually below 1000 ft, and the median is 454 ft. The mean of the distribution is pulled to the right to 569 ft by the skewness. As discussed previously, the mean of the distribution is most often a better prediction than the median in this application. This is due in part to the log-transformation of the data which preserves the inherent skewness of the predictors and predictands. The mean better includes the log-normal skewness of the posterior distribution. Table 35 displays the high density (credible) intervals for 68%, 80% and 95%. While analogous to confidence intervals,

these credible intervals literally correspond to the probability that an observed value will fall within the interval. For example, for this case, there is a 68% chance the observed ceiling height will fall between 124 ft and 653 ft.

Next, using the PPDs, probabilistic forecasts for univariate or multivariate conditions can be calculated. Table 36 displays univariate probabilistic forecast for ceiling height and visibility.

Table 36. Probabilistic Forecasts for 2017091907

	Ceiling (%)	Visibility (%)
LIFR (<500 ft, <1 mile)	.56	.01
IFR (500-1000 ft, 1-3 miles	.28	.16
MVFR (1000-3000 ft, 3-5 miles)	.11	.29
VFR (>3000ft, >5 miles)	.003	.64

Now, comparing all the probabilistic data for the ceiling height forecast, there was an 80% chance that the observed ceiling height would be between 96 ft and 809 ft. The probability of it being less than 500 ft (LIFR) was 56% and between 500 ft and 1000ft (IFR) was 28%. However, there is a good probability (44% chance) that the ceiling will be greater than 500 ft. Given the NAM forecast for the cloud base height was 375 ft and the NPS mean was 569 ft, the probabilistic information suggests that the observed ceiling height will most likely be greater than the NAM forecast of 375 ft, and could be more than 500 ft. Surprisingly, the observed ceiling height of 700 ft is higher than both the NAM and NPS forecasts but does fit into the 80% credible interval. Most importantly, the NPS forecast successfully pushed the ceiling height forecast into IFR conditions instead of LIFR conditions. The NAM forecast would have resulted in LIFR conditions. This is a useful result because different flight categories have drastically different rules and limitations for aviation operations.

The above analysis can be conducted on every forecast to test the probabilistic information. This type of uncertainty communication is of immense value in decision-making, especially for low ceilings or decreased visibility.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSION

A. SYSTEM LIMITATIONS

1. Supervision

The nowcast system is an example of supervised machine learning use in weather forecasting and can be applied to almost any forecasting application. In the ceiling height and visibility application, the nowcast system can post-process raw forecasts to reduce error, improve skill, and generate PPDs to explicitly communicate uncertainty. However, the forecaster or system, if automated, would need to choose the clustering and conditioning method. Improvements are still needed to transition the system to unsupervised learning for full automation.

2. Data and Model Assumptions

Much of the residual error the postprocessed forecast can likely be attributed to the choice of predictor and predictand data and the model assumptions. The predictor data was extracted from a grid point geographically dislocated from the predictand location. Strictly speaking, this is not necessarily an issue if the relationship between the two locations is assumed to be linear. This may not necessarily be true and was not investigated. Moreover, the choice of using NAM NWP data was due to its assumed skill over the continental U.S. Future studies should test the use of different NWP data sets with varying model resolutions as predictors. Next, the use of ASOS observations as predictand data was limited due to artificial caps at 12,200 ft for ceiling and 10 miles for visibility. Future studies should investigate the use of different and more continuous predictand data. Finally, the GLM assumes the posterior distributions are modeled as a normal likelihood function. This is not necessarily wrong but may not be always true for atmospheric phenomena. Future studies should test the use of different likelihood functions.

3. Data Availability

A major goal of this thesis was to demonstrate the ability to generate rigorous and reliable probabilistic forecasts with limited data and short training periods. For military applications, this is a key requirement for usable statistical forecasting systems because much of the world is atmospheric data-sparse, especially for ceiling height and visibility. Future studies should test different training period lengths and seasonal compositing as well as analog forecasting applications in remote locations.

B. SUMMARY

Despite the aforementioned limitations, this study successfully demonstrated the ability to use BE and MCMC methods to construct a nowcasting system that post-processes raw NWP data. The posterior forecasts reduced error, improved forecast skill and communicated useful probabilistic information. The system is computationally efficient with average model run times of about 3-4 minutes on an off-the-shelf MacBook Pro. Similarly, the system requires short training periods a just a few months of data, as opposed to other statistical post-processing systems that require years of data. This is due to large part to the power of the direct application of Bayesian inference to the forecasting problem, simple yet effective machine learning techniques, and MCMC methods. It cannot be overstated that the single most important value in a nowcasting system of this type is the ability to generate posterior predictive distributions that explicitly communicate uncertainty. In the hands of a forecaster or decision-maker, this information can be the deciding factor in the confidence of a forecast or a go/no-go call due to weather. Moreover, this dynamically-forced statistical post-processing system has a wide scope of physical battlespace awareness applications and can be generally applied to most physical phenomena from advanced climate support to ocean acoustics to physical oceanography. Finally, this system is meant to complement, not replace, dynamical forecasting. Both methods generate unique and useful results, and each method has its strengths and weaknesses. The hallmark of a good forecaster is the ability to know when to use one, the other, or both.

LIST OF REFERENCES

AMS, 2012: American Meteorology Society Glossary of Meteorology. Accessed 26 April 2018, http://glossary.ametsoc.org/wiki/Main_Page.

Bankert, R. L., M. Hadjimichael, A. P. Kuciauskas, W. T. Thompson, and K. Richardson, 2004: Remote Cloud Ceiling Assessment Using Data-Mining Methods. *Journal of Applied Meteorology*, **43**, 1929–1946, doi:10.1175/jam2177.1.

Bocchieri, J. R., and H. R. Glahn, 1972: Use of Model Output Statistics for Predicting Ceiling Height. *Mon. Wea. Rev.*, **100**, 869–879, doi:10.1175/1520-0493(1972)100<0869:uomosf>2.3.co;2.

Butler, M., 2005: The impacts of weather forecasts on military operations: a system for conducting quantitative near real time analyses. Thesis, Department of Meteorology, Naval Postgraduate School, 102 pp.

Cantu, R., 2001: The role of weather in Class A naval aviation mishaps FY 90–98. Thesis, Department of Meteorology, Naval Postgraduate School, 104 pp.

Chmielecki, R. M., and A. E. Raftery, 2011: Probabilistic Visibility Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, **139**, 1626– 1636, doi:10.1175/2010mwr3516.1.

Dixon, M., 2004: Short-term forecasting of airport surface visibility using radar and ASOS. Accessed 19 April 2018, https://ams.confex.com/ams/11aram22sls/techprogram/paper_82036.htm.

FAA, 2017: Accident and Incident Data. Accessed 19 April 2018, https://www.faa.gov/data_research/accident_incident/.

Gallaudet, T. 2017: Naval Oceanography Information Warfare Strategy. Accessed 19 April 2018, http://www.navy.mil/submit/display.asp?story_id=98677.

Gelman, A., 2013: *Bayesian Data Analysis*. Chapman & Hall/CRC, 675 pp.

Glahn, H. R., and D. A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, **11**, 1203–1211, doi:10.1175/1520-0450(1972)011<1203:tuomos>2.0.co;2.

Hansen, B., 2007: A Fuzzy Logic–Based Analog Forecasting System for Ceiling and Visibility. *Weather and Forecasting*, **22**, 1319–1330, doi:10.1175/2007waf2006017.1.

Hilliker, J. L., and J. M. Fritsch, 1999: An Observations-Based Statistical System for Warm-Season Hourly Probabilistic Forecasts of Low Ceiling at the San Francisco International Airport. *Journal of Applied Meteorology*, **38**, 1692–1705, doi:10.1175/1520-0450(1999)038<1692:aobssf>2.0.co;2.

JCS, 2018: Joint Publication 3–59: Meteorological and Oceanographic Operations. Accessed 26 April 2018, http://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_59.pdf.

Leclerc, J., and Joslyn S., 2015: The Cry Wolf Effect and Weather-Related Decision Making. *Risk Analysis*, **35**, 385–395, doi:10.1111/risa.12336.

Leyton, S. M., and J. M. Fritsch, 2004: The Impact of High-Frequency Surface Weather Observations on Short-Term Probabilistic Forecasts of Ceiling and Visibility. *Journal of Applied Meteorology*, **43**, 145–156, doi:10.1175/1520-0450(2004)043<0145:tiohsw>2.0.co;2.

Li, C., 2013: The distinction of turbulence from chaos - rough dependence on initial data. Accessed 19 April 2018, https://arxiv.org/pdf/1306.0470.pdf.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, **20**, 130–41.

Lorenz, E. N., 2003: *The Essence of Chaos*. Routledge, London, 227 pp.

Marzban, C., et al., 2006: MOS, Perfect Prog, and Reanalysis. *Monthly Weather Review*, **134**, 657–663.

Mass, C., 2011: Nowcasting, The Next Revolution in Weather Prediction. *Bulletin of the American Meteorological Society*, Submitted,1-23.

Mueller, C., T. Saxen, R. Roberts, J. Wilson, T. Betancourt, S. Dettling, N. Oien, and J. Yee, 2003: NCAR Auto-Nowcast System. *Weather and Forecasting*, **18**, 545–561, doi:10.1175/1520-0434(2003)018<0545:nas>2.0.co;2.

NRC, 2006: *Completing the forecast: characterizing and communicating uncertainty for better decisions using weather and climate forecasts*. National Academies Press, Washington, D.C., 112 pp.

NWS, 2010: Description of the NAM MOS Alphanumeric Message. Accessed 19 April 2018, http://www.nws.noaa.gov/mdl/synop/namcard.php.

Pedregosa, et al., 2011: Scikit-learn: Machine Learning in Python, *JMLR 12*, 2825-2830.

Rudack, D. E., and J. E. Ghirardelli, 2010: A Comparative Verification of Localized Aviation Model Output Statistics Program (LAMP) and Numerical Weather Prediction (NWP) Model Forecasts of Ceiling Height and Visibility. *Weather and Forecasting*, **25**, 1161–1178, doi:10.1175/2010waf2222383.1.

Ryerson, W., 2012: Toward improving short-range fog prediction in data-denied areas using the Air Force Weather Agency mesoscale ensemble. Dissertation, Department of Meteorology, Naval Postgraduate School, 250 pp.

Sun-Tzu, 1964: *The Art of War*. Oxford, Clarendon Press, 197 pp.

U.S. Navy, 2016: A Design for Maintaining Maritime Superiority. Accessed 19 April 2018, http://www.navy.mil/cno/docs/cno_stg.pdf.

U.S. Navy, 2017: Task Force Ocean. Accessed 19 April 2018, https://www.defense.gov/News/Article/Article/1129450/navy-announces- launch-of-task-force-ocean/.

VanderPlas, J., 2013: Frequentism and Bayesianism: A Practical Introduction. Accessed 19 April 2018, http://jakevdp.github.io/blog/2014/03/11/frequentism-and-bayesianism-a-practical-intro/.

Vislocky, R. L., & Fritsch, J. M., 1997: An Automated, Observations-Based System for Short-Term Prediction of Ceiling and Visibility. *Weather and Forecasting*,**12**, 31–43, doi:10.1175/1520-0434(1997)0122.0.co;2.

Wendt, R., 2017: A hierarchical multivariate Bayesian approach to ensemble model output statistics in atmospheric prediction. Dissertation, Department of Meteorology, Naval Postgraduate School, 221 pp.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences*. Oxford: Academic Press, 704 pp.

Zoufaly, S., 2017: Nowcasting Cloud Fields for U.S. Air Force Special Operations. Thesis, Department of Meteorology, Naval Postgraduate School, 111 pp.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California